The Dissertation Committee for Zilong Xie Certifies that this is the approved version of the following dissertation:

Taking Attention Away from the Auditory Modality: Investigations of the Effect on Speech Processing Using Machine Learning

Committee:

Bharath Chandrasekaran, Supervisor

Christopher G. Beevers

Craig A. Champlin

Chang Liu

Taking Attention Away from the Auditory Modality: Investigations of the Effect on Speech Processing Using Machine Learning

by

Zilong Xie

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2018

Dedication

To my wife Yang and our baby Andrew.

Acknowledgements

I would like to express my gratitude to my academic advisor Dr. Bharath Chandrasekaran for providing me the opportunity to learn and grow without worrying about making mistakes in the path of science and research. One of the many great things I deeply appreciate is that he knows what makes me comfortable in my studies and research, but he sometimes would challenge me to step out of my comfort zone. My colleagues and I once came to the idea that one's Ph.D. advisor is like one's academic 'spouse'. Dr. Bharath Chandrasekaran is undoubtedly a perfect 'spouse' I have been fortunate to work with. Further, I am deeply indebted to him for the endless support during this last year of my Ph.D. while I was fighting the 'unexpected blow' to my life.

I would also like to express my gratitude to Han-Goyl Yi and Rachel Reetzke. Han is one of the smartest persons I have ever met. One thing I admire him is that he is excellent at explaining complicated ideas in a simple, easy-to-understand way. He is like a second mentor to me. I learned all the important research skills from him. Rachel Reetzke is a wonderful workmate. Her great personality and passion for research perfectly complement mine in our collaborative research work. I am also indebted to my other lab colleagues, especially Gangyi Feng and Fernando Llanos, for the invaluable help. This dissertation wouldn't be possible without them.

Last but not least, I would like to express my gratitude to my committee Dr. Christopher G. Beevers, Dr. Craig A. Champlin, and Dr. Chang Liu for the guidance and support in the process of the dissertation.

Taking Attention Away from the Auditory Modality: Investigations of the Effect on Speech Processing Using Machine Learning

Zilong Xie, Ph.D.

The University of Texas at Austin, 2018

Supervisor: Bharath Chandrasekaran

Real-world speech processing often takes place in complex multisensory environments. Listeners may need to prioritize sensory inputs from modalities other than audition. Selective attention is thought to be critical in selecting the sensory modality most relevant to the task at hand. Two critical research questions have driven crossmodal attention research thus far: first, how early does crossmodal attention influences processing in the unattended modality? Second, is there a limitation in attentional resources between sensory modalities? Set within the context of this prior work, this dissertation aims to examine the effects of crossmodal attention on *speech processing* when sensory inputs from *vision* are prioritized. In study 1, we demonstrate that modulating visual perceptual load can impact the early sensory representation of linguistically-relevant pitch contours (Mandarin tones), a suprasegmental feature that is critical to the percept of lexical tones. Further, we provide novel evidence that the impact of the visual load is highly dependent on the predictability of the incoming speech stream. In study 2, we utilized ecologically valid, continuous speech and tested the extent to which dividing attention to a visual task affects neural processing of speech signals. We show that dividing attention between auditory and visual tasks leads to both

behavioral and electrophysiological costs in the processing of continuous speech stimuli. The results also demonstrate that the neural encoding of suprasegmental features (e.g., envelope and fundamental frequency) in continuous speech is modulated by diverting attention away from the auditory modality. In contrast, the neural encoding of segmental features (e.g., phonetic features) may be unaffected by taking attention away from the auditory stream. The theoretical and practical implications of the two studies are discussed.

Table of Contents

List of Figures

INTRODUCTION

Real-world speech processing often takes place in complex multisensory environments. Frequently, information from audition and other modalities (e.g. vision) are correlated and complementary. A classic example is audiovisual speech. The human brain integrates audio speech cues and visual speech-reading cues to facilitate speech processing (e.g., Crosse, Butler, & Lalor, 2015; Golumbic, Cogan, Schroeder, & Poeppel, 2013; Sumby & Pollack, 1954; Van Engen, Xie, & Chandrasekaran, 2017; Van Wassenhove, Grant, & Poeppel, 2005; Xie, Yi, & Chandrasekaran, 2014). Extensive research has focused on understanding the effects of multisensory integration on speech processing (see for example Campbell, 2008; van Wassenhove, 2013 for a review). Similarly, in everyday listening situations, it is common that information from audition and other modalities (e.g. vision) are unrelated or even conflicting. For example, we often listen to the radio while driving. Listeners in this situation need to constantly juggle demands across the individual modalities. Selective attention has generally thought to be critical to select the sensory modality most relevant for the task at hand (Spence, 2010). There is ample evidence on selective attention in *unisensory* contexts (i.e., intramodal attention) suggesting that, the allocation of attentional resources to one source (e.g., a distracting talker) exerts detrimental effects on the processing of stimulation from another source (e.g., a speaker of interest) (e.g., Cherry, 1953; Cherry & Taylor, 1954; Ding & Simon, 2012; Mesgarani & Chang, 2012; Power, Foxe, Forde, Reilly, & Lalor, 2012; Woldorff et al., 1993). To date, however, the extent to which taking attention *away* from the auditory modality impacts speech processing is less well understood. Thus, this dissertation aims to examine the effect of crossmodal attention on speech processing when vision is prioritized.

As an introduction, Chapter 1 will provide a review of the literature on crossmodal attention, related theories, and crucial evidence in support of or against the theories. Chapter 2 will then provide an overview of the research goals and proposed studies for the dissertation.

Chapter 1: A review of literature on crossmodal attention research, related theories, and relevant studies

At any waking moment, our brain is bombarded with sensory information from multiple modalities. To navigate in such multisensory environments and fulfill our goals, we often need to select information from one specific sensory modality and filter out information from other sensory modalities. For example, when working on the dissertation in a cafeteria, it is advantageous to focus on one's computer screen and ignore the background noise.

Until now, two critical issues have remained unresolved in the literature involving crossmodal attention. One concerns how early crossmodal attention influences processing in the unattended modality. The other concerns whether there is a limitation in attentional resources between sensory modalities. The extant empirical evidence regarding the two questions are not conclusive. In the following sections, I provide a summary of rivaling conceptualization for each issue and relevant studies.

HOW EARLY IS THE GATING OF NEURAL PROCESSING BY CROSSMODAL ATTENTION?

Selection attention entails the targeted selection of task-relevant information while ignoring or suppressing task-irrelevant information. A fundamental question in prior work examining attention is when (or at what processing stage) the task-irrelevant information is filtered out. For decades, it has been debated whether selective attention takes place at an early or late stage of information processing. Per early-selection theory, due to a limited capacity of perceptual processing, the processing of task-irrelevant (unattended) stimuli can be prevented at an early perceptual stage (e.g., Broadbent, 1952, 1958; Treisman, 1969). In contrast, based on the late-selection theory, there are no limits in the perceptual processing capacity, and task-irrelevant (unattended) stimuli can be

fully processed as task-relevant (attended) stimuli. Attention operates only on later processes such as memory or behavioral response (e.g., Deutsch & Deutsch, 1963; Duncan, 1980; Dux, Ivanoff, Asplund, & Marois, 2006).

To resolve the 'early versus late selection' debate, Lavie and Tsal (1994) proposed that a major factor in determining the locus of selection is perceptual load. The perceptual load theory (see for example Lavie, 2005; Murphy, Groeger, & Greene, 2016 for a review) holds that perception is of limited capacity and proceeds automatically until that capacity is exhausted. Performing a task of high perceptual load may already exhaust the perceptual processing capacity and does not allow the processing of task-irrelevant stimuli. This leads to performance on the task-irrelevant stimuli that are consistent with the early-selection views. However, when performing a task of *low perceptual load*, both task-relevant and -irrelevant stimuli are processed. Hence, late-selection is required to prevent that the task-irrelevant stimuli gain control over behavior. A major factor in determining the efficiency of late selection is the level of load on cognitive control processes (e.g., working memory) imposed by the primary task. Increasing cognitive load can lead to a failure of late selection, and, as a result, enhanced processing of distractors (Lavie, 2010).

The perceptual load theory has received extensive support from studies conducted within sensory modalities (e.g., Lavie, 1995; Torralbo et al., 2016) as well as across modalities (e.g., Ciaramitaro et al., 2017; Klemen et al., 2009; Kreitz, Furley, Simons, & Memmert, 2016; Macdonald & Lavie, 2011; Molloy et al., 2015; Raveh & Lavie, 2015). However, regarding the role of cognitive load in the processing of task-irrelevant (unattended) stimuli, there is behavioral and neuroimaging data to suggest that increasing cognitive load leads to a reduction in the processing of task-irrelevant distractors (e.g., Hadar, Skrzypek, Wingfield, & Ben-David, 2016; Halin et al., 2015; Hunter & Pisoni,

2017; Simon, Tusch, Holcomb, & Daffner, 2016; Sörqvist, Dahlström, Karlsson, & Rönnberg, 2016; Sörqvist, Stenfelt, & Rönnberg, 2012) that may occur early before sensory information reaches the cortex (Sörqvist et al., 2012). These findings contrast with the predictions of the perceptual load theory. For example, Halin, Marsh, and Sörqvist (2015) instructed participants to perform a visual-verbal *n*-back task with two levels of cognitive load (1-back vs. 2-back) while ignoring concurrent auditory stories. In an immediate follow-up memory test, they found poorer memory of the stories in the high cognitive load condition (2-back) than that in the low cognitive load condition (1-back). In an electroencephalography (EEG) study, participants were confronted with a visual-verbal *n*-back task with varying cognitive load (1-, 2-, and 3-back) while ignoring simultaneously presented tone bursts. Participants also completed a condition wherein they actively listened to the tone bursts by monitoring the occurrences of deviant tones. They found that wave V in the auditory brainstem responses (ABRs) evoked by the tone bursts showed reduced magnitude with higher cognitive load (2- and 3-back) relative to low cognitive load (1-back and active listening) (Sörqvist et al., 2012). Wave V of the ABR is thought to reflect early auditory processing primarily originated from the pre-attentive subcortical auditory nuclei (Møller & Jannetta, 1985).

IS THERE A LIMITATION IN ATTENTIONAL RESOURCES BETWEEN MODALITIES?

It is generally argued that there are limits in the amount of information humans can selectively attend to (see Marois & Ivanoff, 2005 for a review). The limitations in attentional selection have been hypothesized as a pool of attentional resources (Kahneman, 1973; Lavie, 2005; Wickens, 1991, 2008). This pool of resources can be allocated to the tasks at hand until it is exhausted. For example, engaging in a task that is not very challenging may consume only a portion of attentional resources, and there are

spare attentional resources that can be allocated to another task simultaneously. But in conditions involving a highly challenging primary task, attentional resources may already be exhausted and little or none would be left for another task.

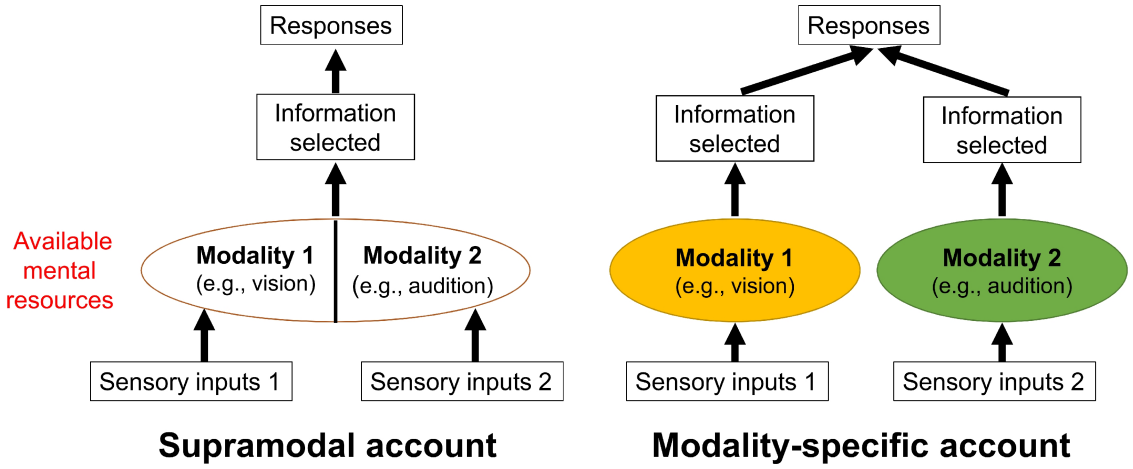

Figure 1: Schematics to illustrate the *supramodal* (left) and *modality-specific* (right) accounts of attentional resources for attentional selection across modalities.

A substantial body of evidence has demonstrated that selection within a modality suffices limits in attentional resources (e.g., Alais, Morrone, & Burr, 2006; Broadbent, 1952; Kastner & Ungerleider, 2001; Neisser & Becklen, 1975; Porcu, Keitel, & Müller, 2014; Torralbo, Kelley, Rees, & Lavie, 2016). Yet, there is still debate about whether similar resource limitations pertain to attentional selection across modalities. Some researchers have argued that there is a central, limited pool of attentional resources that are shared across modalities. The depletion of resources by one modality diminishes available resources in the other modality (Broadbent, 1957; Ciaramitaro, Chow, & Eglington, 2017; Jolicoeur, 1999; Klemen, Büchel, & Rose, 2009; Macdonald & Lavie,

2011; Molloy, Griffiths, Chait, & Lavie, 2015; Raveh & Lavie, 2015). According to the *supramodal* account of attentional resources (left panel in Figure 1), attention to one sensory modality impairs information processing within another modality. Evidence from numerous behavioral (Causse, Imbert, Giraudet, Jouffrais, & Tremblay, 2016; Ciaramitaro et al., 2017; Halin et al., 2015; Hunter & Pisoni, 2017; Macdonald & Lavie, 2011; Mattys, Barden, & Samuel, 2014; Mattys & Wiget, 2011; Mitterer & Mattys, 2017; Murphy & Greene, 2017; Raveh & Lavie, 2015; Sinnett, Costa, & Soto-Faraco, 2006) and neuroimaging (Klemen et al., 2009; Molloy et al., 2015; Shomstein & Yantis, 2004) studies support this notion. For example, Lavie and colleagues instructed participants to detect a tone during the performance of a visual letter search task of high (target similar to distractors) and low (target dissimilar to distractors) perceptual load. They found that high perceptual load is associated with reduced sensitivity in detecting the tones (Macdonald & Lavie, 2011; Molloy et al., 2015; Raveh & Lavie, 2015). Further, they utilized magnetoencephalography (MEG) to measure associated neural responses while participants performed the visual search tasks and ignored concurrent tone stimuli. They found increased early visual cortical evoked activity (~100 ms after tone onset) during high relative to the low visual load condition. Interestingly, this was accompanied by a substantial reduction in early auditory cortical evoked activity (~100 ms after tone onset) to the task-irrelevant auditory stimuli for high than low visual load (Molloy et al., 2015). The reserve pattern of the visual load effects on visual and auditory processing is what the *supramodal* account of attentional resources predicts based on shared, limited attentional resources between modalities.

In contrast to the *supramodal* account of attentional resources, others have argued that each sensory modality individually has a limited pool of attentional resources. These *modality-specific* attentional resources operate independently (Alais et al., 2006; Arrighi,

Lunardi, & Burr, 2011; Duncan, Martens, & Ward, 1997; Keitel, Maess, Schröger, & Müller, 2013; Parks, Hilimire, & Corballis, 2011; Porcu et al., 2014). Based on this *modality-specific* account of attentional resources (right panel in Figure 1), diverting attention to one sensory modality may not directly influence information processing in another modality. Findings from several behavioral studies support this notion (Alais et al., 2006; Arrighi et al., 2011; Duncan et al., 1997; Masutomi, Barascud, Kashino, McDermott, & Chait, 2016). For example, Alais, Morrone, and Burr (2006) confronted participants in a dual-task paradigm. The two tasks were presented either in the same modality (both tasks in the visual or auditory modality) or different modalities (one in the visual modality and the other in the auditory modality). The primary visual task was a contrast discrimination task, i.e., to discriminate which of two grating patches was higher in contrast, and the primary auditory ask was a frequency discrimination task, i.e., to distinguish which of two tones was higher in pitch; the secondary visual task was to detect among a brief central array of dots whether one dot was brighter than the others, and the secondary auditory task was to detect whether a brief triad of tones formed a major or a minor chord. Participants were instructed to treat the two tasks as equally important. They found that performance on the primary tasks was unaffected by a secondary task from the *other* modality. In contrast, performance was significantly deteriorated by a secondary task from within the same modality. Notably, varying the difficulty of the secondary task had very little effect on these patterns of results. Neuroimaging findings further provide evidence for the *modality-specific* account (Chait, Ruff, Griffiths, & McAlpine, 2012; Keitel et al., 2013; Parks et al., 2011; Porcu et al., 2014; Rees, Frith, & Lavie, 2001). For example, Parks, Hilimire, and Corballis (2011) instructed participants to monitor the occurrences of targets in a rapid central stream of visual stimuli of high (targets were identifiable by the conjunction of two features) and

low load (targets were identifiable by a single feature). Participants performed the task alone and in the presence of task-irrelevant visual and auditory distractors. They found that EEG activity evoked by the visual distractors (over occipital electrodes) was significantly reduced during high (relative to low) perceptual load, whereas EEG responses to the auditory distractors (over frontocentral electrodes) remained unchanged with increasing visual load.

Chapter 2: Overview of dissertation goals and proposed studies

STUDY 1: TAKING ATTENTION AWAY FROM THE AUDITORY MODALITY: CONTEXT-DEPENDENT EFFECTS ON THE EARLY SENSORY REPRESENTATION OF SPEECH

In the first study, we examine the extent to which taking attention away from the auditory modality influences the early sensory encoding of speech signals. The scalp-recorded frequency-following response (FFR) (see Figure 2B for examples) provides a noninvasive window into the neural encoding of speech signals at initial stages along the auditory pathway (Chandrasekaran & Kraus, 2010; Krishnan, 2002; Krishnan, Xu, Gandour, & Cariani, 2004; Skoe & Kraus, 2010). The FFR is an electrophysiological response that reflects phase-locked activity to physical properties of acoustic signals (Bidelman, 2015a; Chandrasekaran & Kraus, 2010; Marsh, Worden, & Smith, 1970; Moushegian, Rupert, & Stillman, 1973; Smith, Marsh, & Brown, 1975; Worden & Marsh, 1968). Unlike cortical evoked responses, the frequency of the FFR can be up to about 1000 Hz (Batra, Kuwada, & Maher, 1986; Chandrasekaran & Kraus, 2010). The latency of the FFR typically ranges from 5 to 10 ms (Akhoun et al., 2008; A. King, Hopkins, & Plack, 2016; Skoe, Krizman, Anderson, & Kraus, 2013; Smith et al., 1975), which is earlier than cortical evoked responses (Celesia, Broughton, Rasmussen, & Branch, 1968; Moushegian et al., 1973). The scalp-recorded FFRs derived from electroencephalography (EEG) are hypothesized to reflect activity primarily from subcortical auditory ensembles, i.e., the inferior colliculus (Bidelman, 2015a; Chandrasekaran & Kraus, 2010; Smith et al., 1975).

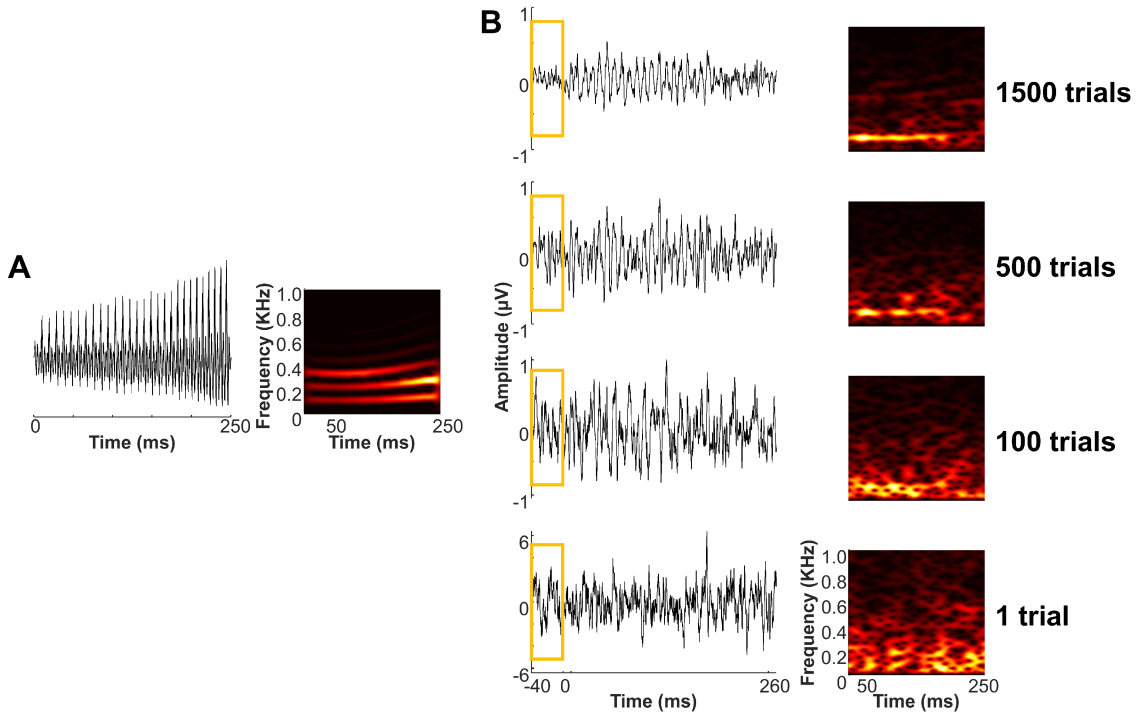

Figure 2: Examples of frequency-following responses (FFRs) elicited to a low-rising (T2) linguistically-relevant pitch pattern. (A) Waveform (left) and spectrogram (right) of the stimulus T2. The frequencies of the fundamental frequency (F0) and higher harmonics (highlighted in red and yellow in the spectrogram; the lowest one represents the F0, and the ones above represent the higher harmonics) increase over time. (B) Waveforms and spectrograms of the FFRs averaged across a different number of trials (from top to bottom): 1500, 500, 100, and 1. Note that, it is difficult to observe any activity pattern in the FFRs that is similar to the stimulus with a smaller number of trials (e.g., 1 trial). The orange rectangles in the waveforms denote responses before the onset of stimulus (-40 to 0 ms), which reveal the level of noise in the FFRs.

The extent to which crossmodal attention modulates FFRs in human listeners is a subject of intense debate. Galbraith and colleagues (2003) demonstrated that visual attention decreased the amplitude of FFRs to repetitive tones. Hairston et al. (2013) also showed that visual attention reduced the robustness of FFRs to repetitive sounds. These

findings are consistent with substantial animal work that attention can modify the early sensory encoding of auditory signals (e.g., Oatman & Anderson, 1980; Oatman & Anderson, 1977; Slee & David, 2015). The top-down modulation of auditory processing is possibly executed via corticofugal pathways, which are feedback projections from primary and association auditory cortical areas to subcortical auditory structures (see Winer, 2005 for a review). In a recent study, Varghese et al. (2015) concluded no effects of visual attention on the human FFR. In their study, participants were instructed to detect target digit sequences in the visual modality while ignoring a concurrent auditory stream of digits. They compared the visual task condition to two active listening conditions: monaural listening, i.e., attending to a monaural digit stream; and selective listening, i.e., attending to one of two streams of digits presented dichotically. They found no significant differences in the robustness of FFRs to auditory digits across the three conditions.

The inconsistencies in the literature describing crossmodal attentional effects in the human FFR need further consideration, given the observation of crossmodal attentional effects using other types of functional measurements of early auditory processing, such as otoacoustic emission (OAE; Srinivasan, Keil, Stratis, Carr, & Smith, 2012; Wittekindt, Kaiser, & Abel, 2014) and auditory brainstem response (ABR; Sörqvist et al., 2012). The mixed findings across studies using the FFR as a metric may reflect task procedures that create large variability in the degree to which listeners need to disengage their attention from the auditory stimuli. For example, as mentioned in Chapter 1, Sörqvist et al. (2012) demonstrated that, the wave V of ABR, a transient counterpart of the FFR, remained unchanged from an active listening condition to a condition where listeners performed a visual task of low attentional demand (i.e., 1-back). A significant reduction in the wave V amplitude was observed only when the attentional demand of the

visual task increased (2- and 3-back). Moreover, task paradigms utilized in previous studies did not require listeners to consistently maintain attention on a simultaneously presented visual task, leaving opportunities for attentional capture by the auditory stimuli. Furthermore, due to the low signal-to-noise ratio of FFR at the single-trial level (see Figure 2B for examples), the extant studies typically rely on FFRs averaged across thousands of trials (Skoe & Kraus, 2010). Animal studies have demonstrated that the hypothesized generator of the FFR, the inferior colliculus, exhibits rapid stimulus-specific adaptation (SSA), i.e., decreased responsiveness to repetitive elements in the signal (Anderson & Malmierca, 2013; Ayala & Malmierca, 2013; Malmierca, Cristaudo, Pérez-González, & Covey, 2009; Pérez-González, Malmierca, & Covey, 2005). Thus, it is likely that an averaged FFR signal across thousands of trials reflects an aggregate of multiple responses that have undergone SSA due to the lack of novelty in the auditory stimuli. Such adaptation may preclude the observation of effects of flexible cognitive demands, such as shifting attention across modalities.

To address these limitations, study 1 adopts the crossmodal attention paradigm (Figure 3B) from Molloy et al. (2015) that effectively modulates behavioral and neural (cortical) responses to auditory stimuli. Also, this paradigm may be advantageous in minimizing sound adaptation effects over previous FFR studies, because: 1) the interstimulus interval is at least about one second; 2) the repetition of each auditory stimulus is less than 100. To facilitate the analysis of FFRs with small number of trials (as opposed to typical studies), we apply a supervised machine learning approach to decode speech signals from FFRs (Figure 6B), in light of recent progress in the development of machine learning approaches to decode FFRs with reduced number of trials (Llanos, Xie, & Chandrasekaran, 2017; Yi, Xie, Reetzke, Dimakis, & Chandrasekaran, 2017). Briefly, the supervised machine learning approach creates a

classifier that separates FFRs to various speech sounds in a training dataset, and then evaluate how well the classifier generalizes to novel examples of FFRs (i.e., testing dataset) (Kotsiantis, Zaharakis, & Pintelas, 2007). The ability of the classifier to generalize to the testing FFR dataset is used to index differences in the early sensory representation across speech stimuli.

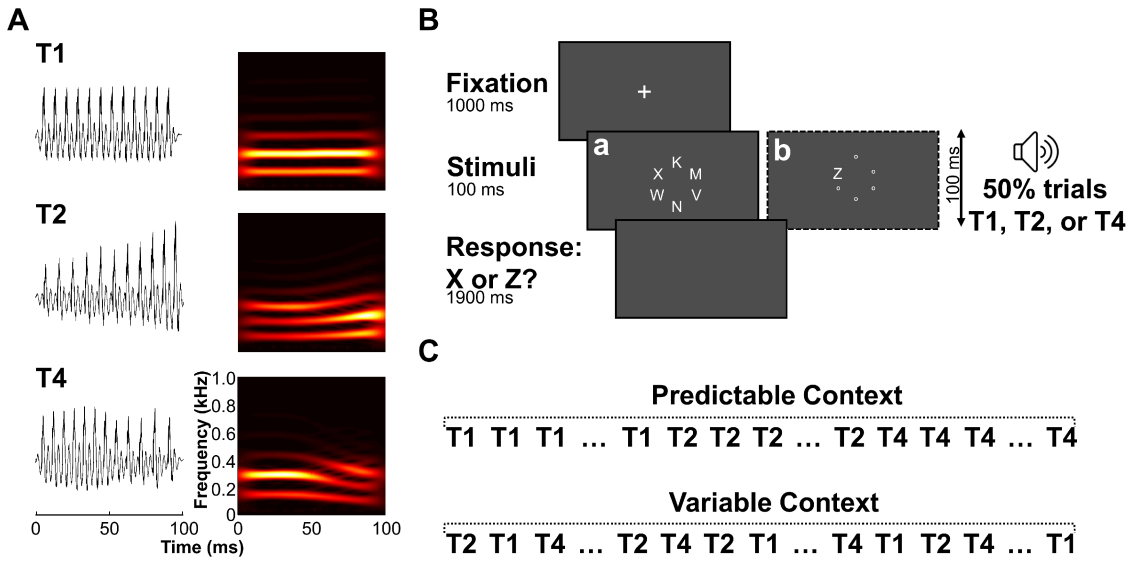

Figure 3: Stimuli and task design. (A) Waveforms and spectrograms of the auditory stimuli, i.e., 100 ms Mandarin tones T1 (high-level), T2 (low-rising), and T4 (high-falling). (B) Trial structure of the visual search task adapted from Molloy et al. (2015). Each trial began with a 1000 ms fixation cross at the center of the screen. Immediately after, a visual letter array of either high (a) or low (b) load was presented for 100 ms. On a random 50% of the trials, a 100 ms auditory stimulus (Mandarin tone T1, T2, or T4) was presented concurrently with the visual stimuli. In the remaining 50% trials, only the visual letter array was presented. After stimulus representation, a blank screen was presented for a maximum of 1900 ms, during which participants were instructed to identify the visual target (X or Z) as quickly and accurately as possible. Once participants made a response, the task moved to the next trial. (C) The auditory stimuli were presented in either predictable (top) or variable (bottom) contexts. In the predictable contexts, the tones were presented in blocks within which each tone was presented repetitively. In the variable contexts, the tones were presented in a random order. In both contexts, an equal number of each of the three Mandarin tones was used.

STUDY 2: DIVIDING ATTENTION TO THE VISUAL MODALITY IMPAIRS THE PROCESSING OF CONTINUOUS SPEECH

In study 2 we examine the extent to which taking attention away from the auditory modality influence the processing of continuous speech. To date, neuroimaging studies on crossmodal attention effects on auditory processing including our study 1, have typically characterized neural responses to a limited set of repetitive, temporally isolated sounds (e.g., Alho, Woods, Algazi, & Näätänen, 1992; Chait et al., 2012; Davis, 1964; Dyson, Alain, & He, 2005; Hackley, Woldorff, & Hillyard, 1990; Haroush, Hochstein, & Deouell, 2010; Karns & Knight, 2009; Molloy et al., 2015; Porcu et al., 2014; Woods, Alho, & Algazi, 1992; Zhang, Chen, Yuan, Zhang, & He, 2006; but see for example Keitel et al., 2013; Parks et al., 2011). This is mainly due to the constraints imposed by non-invasive neurophysiological recordings from humans. Neural responses from non-invasive neuroimaging modalities are susceptible to physiological noise. To overcome the poor signal-to-noise ratio, hundreds of responses to repetitively presented stimuli are averaged together to provide an estimate of the neural response. In naturalistic environments, however, acoustic signals frequently unfold in a nonrepetitive, continuous fashion. This may be particularly true for speech signals, which are usually uninterrupted for seconds, minutes, or even longer. Thus, while research with a limited set of repeated, temporally discrete sounds have contributed to theory, the lack of naturalness in the stimuli may constrain the generalizability of those findings (Ding & Simon, 2012b; Lalor & Foxe, 2010; Lalor, Power, Reilly, & Foxe, 2009). For example, Bonte et al. (2005) suggest that neural responses to speech units (e.g., syllables) embedded in continuous speech are different from that when they were presented in isolation (even though the stimuli are identical). Consequently, research with more naturalistic stimuli is necessary

to complement our understanding of crossmodal attention influence on speech processing.

Study 2 builds on recent progress in neurophysiological studies on human speech processing, and assess the effects of crossmodal attention on speech processing with continuous speech stimuli (e.g, Crosse et al., 2015; Di Liberto & Lalor, 2017; Di Liberto, O'Sullivan, & Lalor, 2015; Ding et al., 2018; Ding, Chatterjee, & Simon, 2014; Ding & Simon, 2013; Fuglsang, Dau, & Hjortkjær, 2017; Khalighinejad, Cruzatto da Silva, & Mesgarani, 2017; Kong, Mullangi, & Ding, 2014; Lalor & Foxe, 2010; Makov et al., 2017; Mirkovic, Debener, Jaeger, & De Vos, 2015; O'sullivan et al., 2014; Power et al., 2012; Presacco, Simon, & Anderson, 2016; Puschmann et al., 2017; Puvvada & Simon, 2017). Specifically, in study 2, participants performed a visuospatial *n*-back task at two levels of demand (0-back, low demand vs. 3-back, high demand) (Figure 4) (Jaeggi et al., 2007) while listening to narrative stories of approximately 60 s long. The visuospatial *n*-back task was administrated as a dual-task, such that participants treated the visual task as the primary task and attended to the stories as a secondary task. In a third condition, participants were presented with similar visual-audio stimuli, but were instructed to attend to the auditory stimuli and ignore the visual stimuli (active listening). Across all three conditions, two multiple-choice comprehension questions for the story segments were asked at the end of each trial to derive a behavioral measure of attention modulation on continuous speech processing. EEG responses to the stories were recorded and compared across the three task conditions (active listening, visuospatial 0- and 3-back), to evaluate the effects of crossmodal attention on the neural processing of speech signals.

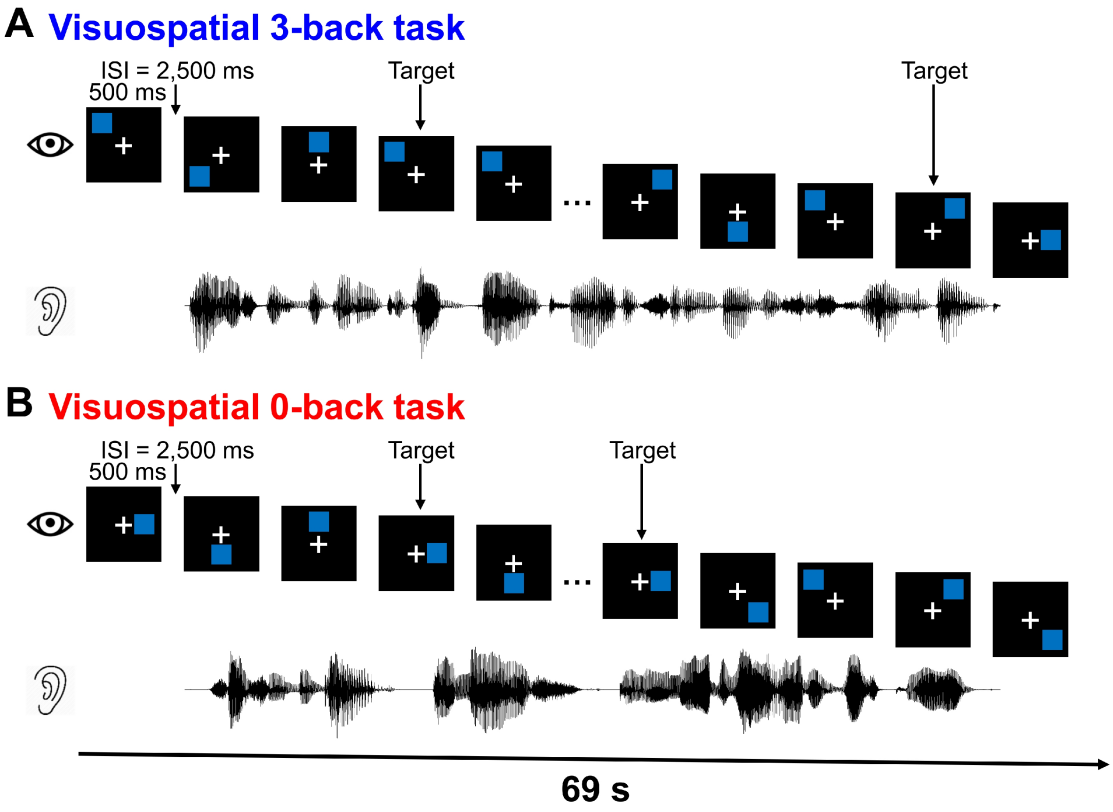

Figure 4. Schematic of the visuospatial *n*-back tasks. Visuospatial stimuli were presented concurrently with a story segment, wherein the story segment began 3 s after the visuospatial stimuli, i.e., starting at the onset of the second blue square in the sequence, and ended earlier than the visuospatial stimuli. Participants' primary task was to respond to the visuospatial stimuli, and a secondary task was to attend to auditory stimulus with whatever they had left. For the 3-back condition, participants responded whether the current blue square matched the one 3 positions back in the sequence (i.e., appearing at the same location as). For the 0-back condition, participants responded whether the current blue square matched a predefined target, which was always the first square in the sequence. Participants responded only to the targets. At the end of the trial, participants were asked two multiple-choice comprehension questions for the story segment. (A) An example trial for the 3-back condition. (B) An example trial for the 0-back condition. ISI: interstimulus interval.

THE CURRENT EXPERIMENTS

Chapter 3: Taking attention away from the auditory modality: Context-dependent effects on the early sensory representation of speech

INTRODUCTION

The central nervous system is constantly extracting meaningful patterns or regularities from the incoming stimuli (Stefanics, Kremláček, & Czigler, 2014; Winkler, Denham, & Nelken, 2009). Sensory systems adjust their response properties based on the stimulation context of the sensory inputs (Nelken & Ulanovsky, 2007). At the same time, we often find ourselves focusing attention on a task in one modality while ignoring information from other sensory modalities. This raises two important questions: to what extent does allocating attentional resources to one modality preclude the processing of task-irrelevant stimuli in another modality, and to what extent does such attentional modulation interact with the processing of regularities of task-irrelevant stimuli? In the current study, we addressed these questions by examining the influence of attending to a visual task on the processing of task-unrelated speech stimuli with varying predictability.

Per the *supramodal* account of attentional resources, visual and auditory processing share capacity-limited neural resources, and the depletion of resources by one modality diminishes available resources to the other modality (Broadbent, 1957; Ciaramitaro et al., 2017; Jolicoeur, 1999; Klemen et al., 2009; Macdonald & Lavie, 2011; Molloy et al., 2015; Raveh & Lavie, 2015). Behavioral studies have provided support for the *supramodal* account of attentional resources, revealing the impact of modulating visual attention on audition. For example, crossmodal studies show that visual tasks of high perceptual load are associated with a reduced detection sensitivity of task-irrelevant auditory stimuli, demonstrating load-induced *inattentional deafness* (Macdonald & Lavie, 2011; Molloy et al., 2015; Raveh & Lavie, 2015). The neural mechanisms underlying

these modulatory influences are still under debate. From late-selection views, while incoming sensory inputs are fully processed by both modalities at perceptual levels, the sensory information that gains prominence is determined at a more central level of processing (e.g., Deutsch & Deutsch, 1963; Duncan, 1980; Dux, Ivanoff, Asplund, & Marois, 2006). However, a recent MEG study revealed a substantial reduction in early auditory cortical evoked activity to task-irrelevant auditory stimuli during high, relative to low, visual load (Molloy et al., 2015). Sörqvist et al. (2012) demonstrated that the magnitude of ABR, which reflects early auditory subcortical processing, significantly reduced with increasing visual task demands. These findings suggest that crossmodal influences can potentially be discerned even during the *encoding* stage of auditory processing, consistent with the early-selection views (e.g., Broadbent, 1952, 1958; Treisman, 1969). Here, we directly test the hypothesis that visual attention can influence the early sensory representation of speech signals by examining the frequency-following response (FFR), a scalp-recorded 'neurophonic' component that faithfully captures phase-locked activity to periodic stimuli (Bidelman, 2015a; Chandrasekaran & Kraus, 2010; Marsh et al., 1970; Moushegian et al., 1973; Smith et al., 1975; Worden & Marsh, 1968).

FFRs have been extensively used to index the fidelity of early, pre-attentive speech encoding in humans (Chandrasekaran & Kraus, 2010; Krishnan, 2002; Krishnan et al., 2004; Skoe & Kraus, 2010). The scalp-recorded FFRs derived from electroencephalography (EEG) are hypothesized to reflect activity primarily from subcortical auditory ensembles (Bidelman, 2015a; Chandrasekaran & Kraus, 2010; Smith et al., 1975). However, there may also be cortical contributions to the FFR as well (Coffey, Herholz, Chepesiuk, Baillet, & Zatorre, 2016). A recent proposal suggests that the FFRs reflect an integrated, dynamic interplay between pre-attentive cortical and

subcortical circuitry (Kraus & White-Schwoch, 2015). This interplay is facilitated by ascending, as well as corticofugal pathways (King & Bajo, 2013; Malmierca, Anderson, & Antunes, 2015; Suga, 2008).

Prior work has demonstrated that early sensory representations of speech signals, as indexed by FFRs, are highly sensitive to stimulus context. Specifically, evidence has shown that sensory fidelity is enhanced for speech sounds presented in predictive contexts relative to less predictable contexts (Chandrasekaran, Hornickel, Skoe, Nicol, & Kraus, 2009; Lau, Wong, & Chandrasekaran, 2016; Lehmann, Arias, & Schönwiesner, 2016; Parbery-Clark, Strait, & Kraus, 2011; Slabu, Grimm, & Escera, 2012). The neural mechanism underlying context-dependent modulation of early speech representation is unclear. Two distinct mechanisms may be at play: 1.) corticofugal modulation that selectively enhances the representation of regularities in the signal via predictive processing; and 2.) novelty detection, which reflects predominantly local stimulus-specific adaptation (SSA) (Chandrasekaran, Skoe, & Kraus, 2014) that enhances the representation of novel elements in the signal (Anderson & Malmierca, 2013; Ayala & Malmierca, 2013; Malmierca et al., 2009; Pérez-González et al., 2005).

In the current study, we examined the impact of manipulating visual perceptual load on the FFRs to speech signals under predictable and less predictable contexts. Native Chinese speakers performed a visual search task of high or low perceptual load (Figure 3B). On a random subset of trials (50%), task-irrelevant Mandarin lexical tones (Figure 3A) were presented in a predictable context or a variable context (Figure 3C). We recorded early cortical evoked activity, along with FFRs to the tones, and utilized a machine learning approach to decode speech category (Mandarin tone) information from the FFRs. We evaluate the extent to which decoding performance, which reflects the fidelity of stimulus encoding, is modulated by visual load and stimulus context. Results

reveal that, when the speech stimuli were presented in predictable contexts, the decodability of FFRs was *reduced* during higher visual load. But when irrelevant speech stimuli were presented in variable contexts, increasing visual load *increased* the decodability of FFRs. We propose that a demanding visual task takes resources away from the auditory cortex, which 'releases' control from online predictive processes. Under these conditions, we posit that stimulus encoding, as indexed by the FFRs, is geared towards the processing of less predictable (more novel) events. In contrast, in a less demanding visual task, the auditory cortex has available resources to enhance sensory tuning via predictive processing.

METHODS

Participants

Twenty adult native speakers of Mandarin Chinese (9 females; 19 to 35 years old) took part in the study. All participants self-reported no previous history of hearing problems or neurological disorders. Participants underwent audiometric testing to ensure pure-tone thresholds ≤ 25 dB hearing level (HL) for octaves from 250 to 4000 Hz (less than 15 dB difference between the two ears) and had normal or corrected to normal vision. Each participant provided written, informed consent and received monetary compensation for their participation. The experimental protocol was approved by the Institutional Review Board at The University of Texas at Austin.

Stimuli and Apparatus

Participants completed a visual search task in an acoustically shielded booth. The visual search is similar to the task described in Experiment 1 in Molloy et al. (2015). The visual stimuli were presented on a zero latency VIEWPixx/EEG scanning LED-backlight LCD monitor (height: 29.1 cm, width: 52.2 cm; display resolution: 1920*1080; refresh

rate: 120 Hz), placed ~100 cm from the participants' eyes. As shown in Figure 3B, the stimuli for the visual task consisted of six letters spaced about equally (subtending a viewing angle of ~1.5°) around the center of the screen. The letters and the fixation cross were presented in white, and the display background was dark gray (RGB values: 77, 77, 77). One of the six letters was the target letter, X or Z (size = 0.55 × 0.45°) that occurred in equal probability. In the high load condition (display a in Figure 3B), letters K, W, V, N, and M (all with the same size as the target letters) served as the distracting items. In the low load condition (display b in Figure 3B), five smaller O's (size = 0.19 × 0.15°) were the distracting items. On each trial, we randomized the positions of the letters so that there was an equal probability for the target letter to appear in each of the six positions.

On a random 50% of trials, auditory stimuli were presented concurrently with the visual letter array (Figure 3B) via insert earphones (ER-3; Etymotic Research, Elk Grove Village, IL) at 60 dB sound pressure level (SPL). The SPL levels were measured by presenting sounds via the same insert earphones for experiment to a Brüel & Kjaer artificial ear (type 4152) connected with a Brüel & Kjaer hand-held analyzer (type 2250-L). The auditory stimuli were 100-ms-long, diotically presented, linguistically-relevant pitch patterns (Mandarin tones): T1 (high-level), T2 (low-rising), and T4 (high-falling) (see Figure 3A for the waveforms and spectrograms). The three tones differ in fundamental frequency (F0) contours: T1 has a relatively flat F0 contour, T2 has a rising F0 contour, and T4 has a falling F0 contour. F0 contour is the primary acoustic cue for native Mandarin listeners to differentiate tones (Gandour, 1983; Howie, 1976). The tones were composed of the same syllable /i/ and were produced by a male native speaker of Mandarin Chinese, recorded at a sampling rate of 44.1 kHz. In pilot testing, native

speakers (n = 5) reliably identified the tone categories with a high degree of accuracy (> 90%).

Design and Procedure

Participants completed the visual task in conditions of either high (display a in Figure 3B) or low (display b in Figure 3B) visual load, wherein the auditory stimuli were presented in either a predictable or a variable context. As shown in Figure 3C, in the predictable context (top), the tones were presented in blocks within which each tone was presented repetitively. In the variable context (bottom), the tones were presented in a random order. In both contexts, an equal number (96) of each of the three tones were used. Therefore, our study consisted of a 2 (visual load: high vs. low) × 2 (stimulus context: predictable vs. variable) within-subject design. The four experimental conditions were divided into two sessions that were separated by seven to twelve days for 17 of 20 participants. For the remaining three participants, the session intervals were between 91 to 108 days due to scheduling conflicts. In each session, the auditory stimuli were presented in only one of the stimulus contexts (predictable or variable). Half the participants completed the session with predictable contexts first, and the other half completed the session with variable contexts first. In each session, there were 16 blocks (8 low visual load and 8 high visual load) with 72 trials per block, each lasting about 3 minutes. Of the 72 trials in each block, 36 included auditory tones (12 per tone). The two visual load conditions were presented in alternating order, and the order of the two load conditions was counterbalanced across participants.

The experiment was controlled with E-Prime 2.0.10 (Schneider, Eschman, & Zuccolotto, 2002). At the beginning of the study, participants were instructed that they may hear some sounds during the experiment. They were told to ignore the sounds and

focus their attention on the visual task. Participants self-initiated each block. As illustrated in Figure 3B, each trial began with a 1000 ms fixation cross at the center of the screen. Next, a visual letter array of either high (a) or low (b) load was presented for 100 ms. On 50% of the trials, a 100 ms Mandarin tone (T1, T2, or T4) was presented simultaneously with the visual display. In the remaining 50% of trials, only the visual letter array was presented. Immediately after the visual letter array, a blank screen was presented, during which participants were required to identify the visual target as quickly and accurately as possible by pressing designated buttons with their right hand. After their response, the experiment immediately moved to the next trial. Participants had at most 1900 ms to respond. At the end of each block, feedback about accuracy on the visual task was provided to encourage engagement. Between blocks, participants were allowed to take breaks when needed.

Electrophysiological Data Acquisition and Preprocessing

Electrophysiological responses were continuously recorded with Ag/AgCl scalp electrodes placed at high forehead at the hairline (~Fpz; active) referenced to linked mastoids (A1/A2), with another electrode on the mid-forehead as ground. Contact impedance was less than 5 kΩ for all electrodes. Responses were acquired at a sampling rate of 25 kHz using BrainVision PyCorder 1.0.7 (Brain Products, Gilching, Germany). The continuous EEG recordings were differentially bandpass filtered off-line from 1 to 30 Hz (12 dB/octave, zero phase-shift) and from 80 to 2500 Hz (12 dB/octave, zero phase-shift) to predominantly highlight cortical and subcortical sustained auditory electrophysiological responses, respectively (e.g., Bidelman & Alain, 2015; Musacchia, Strait, & Kraus, 2008). The EEG recordings were epoched into segments that are time-locked to the auditory stimuli (cortical ERP: -100 to 300 ms; subcortical FFR: -40 to 150

ms), and to the visual stimuli in visual "alone" trials (cortical ERP: -100 to 300 ms). After baseline correcting each response to the mean voltage of the pre-stimulus region, trials with amplitudes exceeding a pre-defined range (cortical ERP: ± 100 μV; subcortical FFR: ± 50 μV) were rejected.

For auditory cortical ERPs, the artifact-free trials were averaged across all the three tones (T1, T2, and T4) for each condition, and downsampled from 25 kHz to 200 Hz. On average, at least 273.1 (*SD* = 33.34) out of the 288 possible trials (12 trials * 8 blocks * 3 tones) were used in each condition. The grand-average auditory cortical ERPs across the four experimental conditions are shown in Figure 5A. For visual cortical ERPs, the artifact-free trials were averaged for each condition and downsampled from 25 kHz to 200 Hz. On average, at least 287.25 (*SD* = 1.45) out of the 288 possible trials (36 trials * 8 blocks) were used in each condition. The grand-average visual cortical ERPs across the four experimental conditions are shown in Figure 5C.

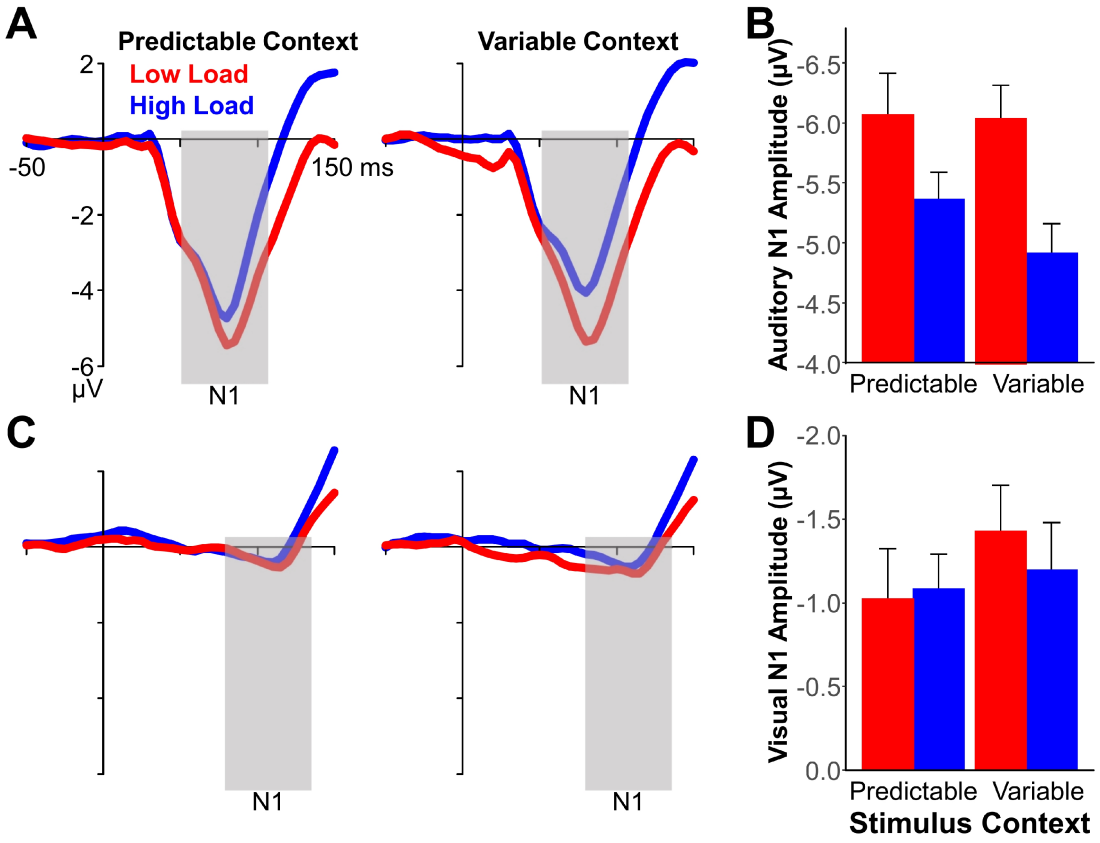

Figure 5: Grand-average cortical evoked responses to the auditory stimuli (A and B), and to the visual stimuli in the visual "alone" trials (C and D) under low (red) or high (blue) visual perceptual load at predictable contexts and variable contexts. (A) Grand-average cortical evoked responses to the auditory stimuli. (B) Mean amplitude of N1 component for the auditory cortical evoked responses. (C) Grand-average cortical evoked responses to the visual stimuli in the visual "alone" trials. The scales of x- and y-axis are the same as in panel A. (D) Mean amplitude of N1 component for the visual cortical evoked responses. The shaded areas in (A) and (C) indicate the time window to find the N1 peak amplitude, which was defined as a 60 ms time window around the N1 component in the grand-average response across the four conditions. Errors bars denote one standard error above the mean.

To capture the FFRs, the artifact-free trials were averaged to produce a sample response to each tone at each condition and downsampled from 25 to 5 kHz. On average, at least 95.10 (*SD* = 1.07) out of the 96 possible trials (12 trials * 8 blocks) were used for any of the tones. Figure 6A displays the grand-average subcortical FFRs to T2 across the four experimental conditions. On average, the signal-to-noise ratio (SNR), computed as the ratio of the root-mean-square amplitude of the post-stimulus region (10 to 110 ms) to that of the pre-stimulus region (-40 to 0 ms), is no lower than 1.24 (*SD* = 0.196) for any of the tones. The SNR was not significantly different across tone, visual load or stimulus context (all *p-value*s > 0.129).

Analysis of Cortical ERPs

Peak amplitude was measured for the N1 component of the auditory and visual cortical ERPs. The auditory N1 component reflects activity generated in auditory cortex and indexes early cortical processing of sounds (Näätänen & Picton, 1987). A prior study showed that the auditory M100, the magnetic equivalent of auditory N1, is reduced during high visual load, relative to low visual load (Molloy et al., 2015). Consistent with this study, we assessed the influence of visual load on auditory N1 responses. Note that, in that study, visual load also modulated visual magnetic response M100. The EEG counterpart of visual M100 is visual P1 (Tobimatsu & Celesia, 2006). However, as displayed in Figure 5C, we did not observe salient visual P1 component in the visual cortical ERPs, possibly due to that the recording site ~Fpz is not optimal to pick up electrophysiological activity related to visual P1 responses (e.g., Alho, Woods, & Algazi, 1994; Vogel & Luck, 2000). We analyzed the visual N1 component in the visual cortical ERPs, with the intent to demonstrate different visual load effect on the visual N1 response as opposed to the auditory N1 response. We can thus make the inference that the

auditory N1 responses in our data primarily reflected auditory activity. As illustrated in Figure 5A and 5C, at each condition, the N1 peak amplitude was taken as the maximal negative amplitude, in a 60 ms time window around the N1 component in the grand-average response across the four conditions. The search for the 60 ms time window was conducted separately for auditory and visual cortical ERPs. The analysis was performed with custom MATLAB scripts (The MathWorks, Natick, MA).

Analysis of FFRs: Decoding Information Related to the Mandarin Tones

Classification analysis was employed to examine the extent to which FFRs evoked by Mandarin tones contain relevant information to discriminate the tones in each of the four experimental conditions. We used a supervised machine learning algorithm (linear support vector machine, linear SVM; Cristianini & Shawe-Taylor, 2000), implemented using the Scikit-learn library (Pedregosa et al., 2011) in python (http://scikit-learn.org/stable/). The linear SVM uses a "one-against-one" approach (Knerr, Personnaz, & Dreyfus, 1990). Specifically, as there were three tones (T1, T2, and T4) in our experiment, linear SVM constructed three classifiers to test the FFR data from all the pairwise combinations of the three tones. The tone label with the highest probability was taken as the classified label. To ensure consistency across experimental conditions, we set the regulation parameter C at a fixed value of 0.1, while keeping other parameters at default values. The selection of this C value was based on our preliminary analysis with grid search to find the best parameter that maximizes tone classification performance.

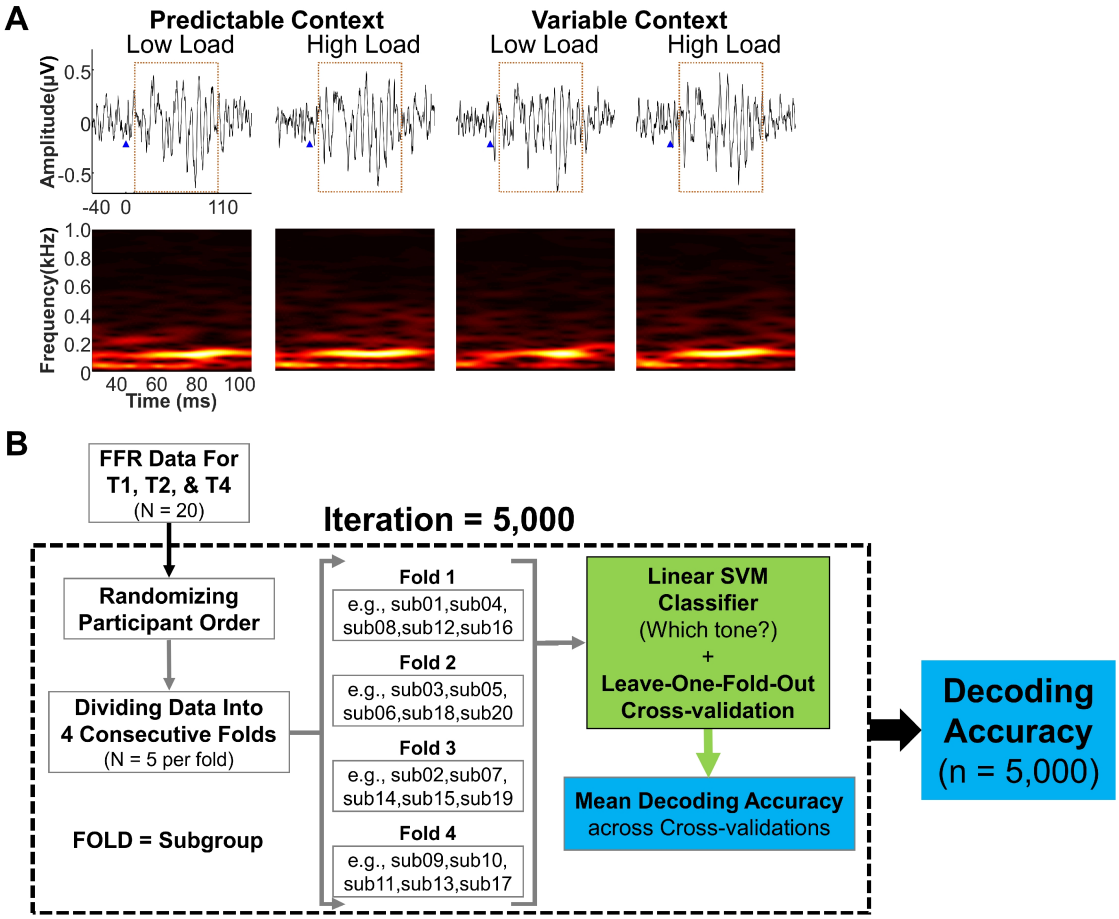

Figure 6. (A) Waveforms and spectrograms of the grand-average FFRs to Mandarin tone T2 (low-rising) across the four experimental conditions. In the waveform plots, the blue triangles indicate the onset of the auditory stimulus. The dashed orange rectangles highlight FFRs from 10 to 110 ms (after stimuli onset) that were used as feature inputs for tone classification analysis. Specifically, the amplitude values (500 values) from this range were used for classification. The spectrograms correspond to FFRs at this range (i.e., 10 to 110 ms). Further, to evaluate the frequency-specific contribution to tone classification, we also applied two bandpass filters (80-180 Hz and 180-600 Hz) to FFRs at range (i.e., 10 to 110 ms) to derive two new types of feature input: 80-180 Hz and 180-600 Hz. Amplitude values (500 values) from the two frequency bands were used for classification analysis, respectively. The frequency band 80-180 Hz covers F0 range of all the three tones (~100 to ~140 Hz). The frequency band 180-600 Hz encompasses the second through fourth harmonics (H2-H4) of all the tones. Note that, as shown in the spectrograms, much of the spectral energy in the FFR concentrates at the frequency range of 80-180 Hz, while limited spectral energy in the FFR was present at the frequency band of 180-600 Hz. (B) Cross-validation procedures for the analysis to decode information related to the Mandarin tones from FFRs.

Cross-validation strategy

To objectively evaluate the performance of the classifier, we used a 4-fold cross-validation strategy with 5,000 iterations. The cross-validation procedures are illustrated in Figure 6B. Each iteration began by randomizing the order of participants. Then, the FFR data were divided into four consecutive folds, with data from five unique participants in each fold. We trained the classifier with three of four folds (i.e., 15 out of 20 participants) and tested whether this training can generalize to the hold-out fold (i.e., the remaining 5 participants). We repeated this cross-validation four times until all the four folds had been tested against. The accuracy of each cross-validation was calculated, which reflected the percentage that the classifier correctly identified the tone labels of the FFR data. At each iteration, we computed the decoding accuracy as the average accuracy across the four cross-validations. We estimated the decoding accuracy for each of the four experimental conditions at each iteration. Hence, 5,000 decoding accuracy values were obtained to evaluate the classifier's performance for each of the four experimental conditions.

Feature selection approaches

In line with our previous study (Xie, Reetzke, & Chandrasekaran, 2017), to account for onset delay reflecting subcortical (specifically midbrain) processes, we selected a region encompassing 10 to 110 ms (after stimulus onset) from each sample response, as representations of FFRs. For the first type of feature input, we used the raw amplitude value at each time point in the FFRs (10 to 110 ms post-stimulus; sampling rate of 5 kHz) (highlighted with orange rectangles in Figure 6A). In other words, the feature input consisted of 500 amplitude values from the FFRs. This type of feature input spans a frequency range of 80 to 2500 Hz. Next, to evaluate the frequency-specific contribution to tone classification, we extracted spectrotemporal information spanning a narrow frequency band (80-180 Hz) that covers F0 range of all the three tones (~100 to

~140 Hz). We chose this frequency band because, as illustrated in the spectrograms of Figure 6A, much of the spectral energy in the FFR is concentrated in the range of F0. We contrasted tone classification based on information from this frequency band (80-180 Hz) with that based on a higher frequency band (180-600 Hz). The higher frequency band encompasses the second through fourth harmonics (H2-H4) of all the tones. To derive these two frequency bands, we applied bandpass filtering (80-180 Hz and 180-600 Hz) to the original FFRs.

Statistical analysis

We applied the following analyses to the decoding performance of the three types of feature inputs separately. In the first analysis, we examined whether the obtained decoding accuracies were significantly above chance. To this end, we applied permutation tests (n = 5,000) to test FFR decoding accuracies against a distribution of decoding accuracies obtained from randomly assigning the labels to the training data (i.e. null distribution). We first estimated the median of the 5,000 decoding accuracies. We then estimated the *p* value using the formula: $p = (a+1)/(n+1)$ (Phipson & Smyth, 2010), where a is the number of decoding accuracies from the null distribution that exceeds the median of the FFR decoding accuracies, and n is the total number of decoding accuracies from the null distribution (i.e., 5,000).

In the second analysis, we examined the effects of visual load and stimulus context on the FFR decoding accuracies. To test the interaction between visual load and stimulus context, we constructed a distribution of the difference between low and high visual load at each stimulus context (i.e., predictable or variable context). This was achieved by calculating the difference in decoding accuracy between low and high visual load at each iteration, resulting in 5,000 difference accuracy scores. We then estimated

the median from the context condition with higher value and tested it against the distribution of difference scores from the context condition with lower median value. We estimated the *p-value*s using the same procedures as described in the first analysis. If a significant interaction was found, we constructed four pairwise comparisons across the four experimental conditions. For each comparison, we estimated the median decoding accuracy from the condition with higher median value and tested it against the distribution of decoding accuracies from the condition with lower median value. We estimated the *p-value*s using the same procedures as described in the first analysis. If no significant interaction between visual load and stimulus context was found, we concatenated the decoding accuracies belonging to the same visual load condition or stimulus context condition and estimated the main effects of visual load and stimulus context using the pairwise comparison method as descried above.

Analysis of FFRs: Tracking of F0 Contours in the Mandarin Tones

To further understand the effects of visual load and stimulus context on the neural representation of the speech stimuli (Mandarin tones) as reflected by the FFRs, we adopted the traditional approach to evaluate the fidelity of neural tracking of F0 contour in the Mandarin tones (e.g., Krishnan, Xu, Gandour, & Cariani, 2005; Krishnan et al., 2004; Xie et al., 2017). Full details of the F0 tracking analysis are described in our previous study (Xie et al., 2017). We modified the parameters of this analysis to optimize the application to the current study. Note that, due to the constraints of using a behavioral task, the FFRs in the current study were averaged across far less number of trials (~95 trials) relative to prior work with similar analysis (several hundreds to thousands; e.g., Krishnan et al., 2005, 2004; Xie et al., 2014). Hence, our results, compared to previous

studies, are less robust to the influence of different sources of noise that may affect FFRs (Skoe & Kraus, 2010).

Extraction of F0 contours

We extracted the F0 contour from the FFRs (10 to 110 ms post-stimulus) using a sliding window (window size = 40 ms, step size = 1 ms) autocorrelation-based procedure (Boersma, 1993). The 40-ms slide window was applied to the time course of each FFR, producing a total of 60 overlapping bins. The autocorrelation function was applied each of the 60 bins to find the maximum (peak) autocorrelation value over a lag value of (1/180 to 1/80 ms), a range that encompasses the periods of the F0 contours for the three Mandarin tones. The peak autocorrelation value, as well as the corresponding lag were recorded for each bin. The frequency of F0 at each bin was taken as the reciprocal of the lag at peak autocorrelation, resulting in a 60-point F0 contour. The same sliding window autocorrelation algorithm was applied to the evoking Mandarin tones to derive the respective stimulus F0 contour.

Evaluation of F0 tracking accuracy

We calculated two metrics to evaluate the robustness of the neural encoding of F0 contour as reflected by the FFRs: stimulus-to-response correlation and peak autocorrelation (e.g., Krishnan et al., 2005; Xie et al., 2017). Details for calculating the two metrics can be found at our previous study (Xie et al., 2017). In short, the stimulus-to-response correlation metric (ranging from 0 to 1) was computed as the normalized cross-correlation between F0 contours between the FFRs and the evoking stimulus. The peak autocorrelation metric (ranging from -1 to 1) was computed as the mean of the peak autocorrelation values across the 60 bins in the FFRs.

RESULTS

Behavioral: Performance on the Visual Search Task

First, we employed a two-way repeated measures analysis of variance (ANOVA) to test the effects of visual load and stimulus context on the performance in the visual search task. For accuracy rate, we found a significant main effect of visual load [$F(1,19) = 65.833$, $p = 1.36 \times 10^{-7}$, $\eta_p^2 = 0.776$], indicating that the mean accuracy was lower in the high load (mean = 90%, $SD = 6.76$) relative to the low load condition (mean = 97.8%, $SD = 2.62$). The main effect of stimulus context did not reach statistical significance [$F(1,19) = 0.123$, $p = 0.729$, $\eta_p^2 = 0.006$]. The interaction between visual load and stimulus context was also not significant [$F(1,19) = 0.145$, $p = 0.707$, $\eta_p^2 = 0.008$]. For task reaction time, we found a significant main effect of visual load [$F(1,19) = 273.7$, $p = 9.71 \times 10^{-13}$, $\eta_p^2 = 0.249$], indicating that mean reaction time increased in the high load (mean = 583.06 ms, $SD = 79.18$) relative to the low load condition (mean = 430.436 ms, $SD = 54.86$). The main effect of stimulus context did not reach statistical significance [$F(1,19) = 3.005$, $p = 0.099$, $\eta_p^2 = 0.137$]. The interaction between visual load and stimulus context was not significant [$F(1,19) = 0.092$, $p = 0.765$, $\eta_p^2 = 0.005$].

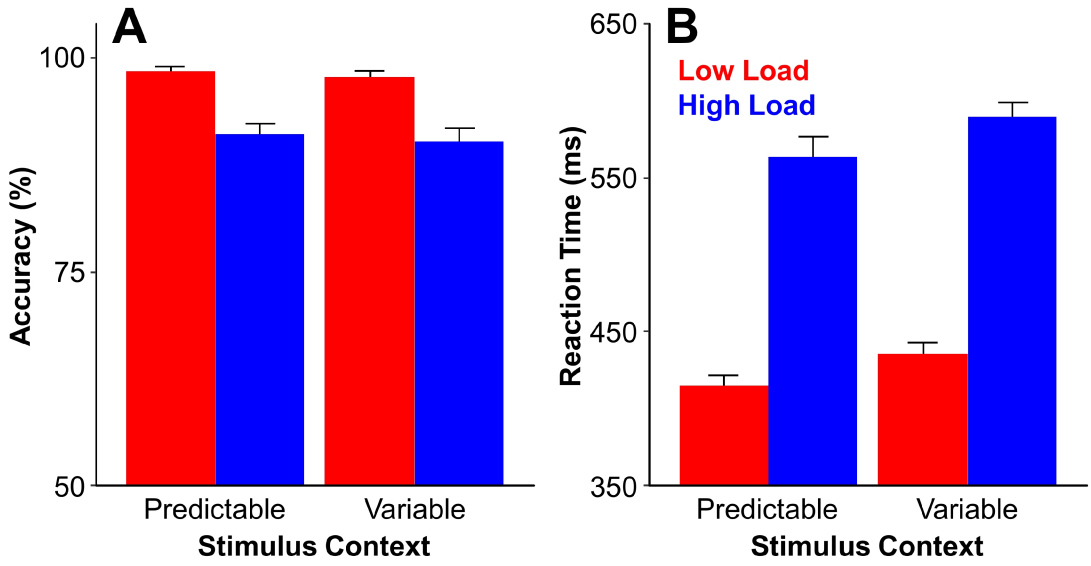

Figure 7. Mean accurate rate (A) and reaction time (B) for identifying the target (X or Z) in the visual search task. Data were presented only for trials with auditory stimuli. The auditory stimuli (Mandarin tone T1, T2, or T4) were presented in predictable or variable contexts. In each stimulus context, the visual stimuli were of low (red) or high (blue) visual perceptual load. Error bars denote 95% confidence intervals.

Next, we tested the effects of visual load and stimulus context on performance in the visual search task by only focusing on the trials when the auditory stimuli were presented concurrently with the visual stimuli. Figure 7 displays the accuracy rate and reaction time. For accuracy rate, we found a significant main effect of visual load [$F(1,19) = 54.956$, $p = 5.097 \times 10^{-7}$, $\eta_p^2 = 0.743$], indicating that the mean accuracy decreased in the high load (mean = 90.68%, $SD = 6.92$) relative to the low load condition (mean = 98.13 %, $SD = 2.36$) (Figure 7A). The main effect of stimulus context did not reach statistical significance [$F(1,19) = 0.54$, $p = 0.471$, $\eta_p^2 = 0.028$]. The interaction between visual load and stimulus context was not significant [$F(1,19) = 0.019$, $p = 0.891$, $\eta_p^2 = 0.001$]. For task reaction time, we found a significant main effect of visual load

[$F(1,19) = 261.46$, $p = 1.46 \times 10^{-12}$, $\eta_p^2 = 0.932$], indicating that mean reaction time increased in the high load (mean = 576.36 ms, *SD* = 82.13) relative to the low load condition (mean = 425.2 ms, *SD* = 54.95) (Figure 7B). The main effect of stimulus context was also significant [$F(1,19) = 7.144$, $p = 0.015$, $\eta_p^2 = 0.273$], indicating slower reaction time in the variable contexts (mean = 512.43 ms, *SD* = 98.92) than the predictable contexts (mean = 489.13 ms, *SD* = 106.47) (Figure 7B). The interaction between visual load and stimulus context did not reach statistical significance [$F(1,19) = 0.127$, $p = 0.726$, $\eta_p^2 = 0.007$]. These findings suggest that when auditory stimuli were presented, a predictable auditory context facilitated performance on the visual task (i.e., faster reaction time) irrespective of the load of the visual task.

Auditory and Visual Cortical ERPs: N1 Amplitude

In line with Molloy et al. (2015), we examined the effects of visual load and stimulus context on the N1 amplitude from the cortical responses to the Mandarin tones, using a two-way repeated measures ANOVA. The mean N1 amplitude of the auditory cortical EPRs are shown in Figure 5B. We found a significant main effect of visual load [$F(1,19) = 9.946$, $p = 5.228 \times 10^{-3}$, $\eta_p^2 = 0.344$], indicating that the mean N1 amplitude decreased in the high load (mean = -5.139 μV, *SD* = 2.836) relative to the low load condition (mean = -6.054 μV, *SD* = 3.345). The main effect of stimulus context did not reach statistical significance [$F(1,19) = 0.537$, $p = 0.473$, $\eta_p^2 = 0.0275$]. The interaction between visual load and stimulus context was not significant [$F(1,19) = 1.169$, $p = 0.293$, $\eta_p^2 = 0.058$].

Note that the Mandarin tones were presented concurrently with visual stimuli with different physical properties at the two load conditions (see Figure 3B). Hence, one possibility is that the load-related differences in N1 amplitude of the cortical response to

Mandarin tones predominantly reflect differences in the visual evoked responses. Such possibility can be refuted because, based on Molloy et al. (2015), we would predict increased cortical responsivity for high visual load relative to low visual load for the visual evoked responses. Further, we directly examined the effects of visual load and stimulus context on the N1 amplitude from visual cortical responses in trials that included only the visual stimuli (i.e., visual "alone" trials). The mean N1 amplitude of visual cortical EPRs is shown in Figure 5D. We did not find significant main effect of visual load [$F(1,19) = 0.153$, $p = 0.7$, $\eta_p^2 = 0.008$] or stimulus context [$F(1,19) = 0.455$, $p = 0.508$, $\eta_p^2 = 0.023$], or significant interaction between visual load and stimulus context [$F(1,19) = 0.861$, $p = 0.365$, $\eta_p^2 = 0.043$]. These results again suggest that the load-related differences in N1 amplitude of the auditory cortical response predominantly reflect differences in auditory activity.

FFRs: Decoding Information Related to the Mandarin Tones

Feature input of 80-2500 Hz

In this analysis, the feature inputs from the FFRs span a frequency range from 80 to 2500 Hz that was used to highlight the FFRs (subcortical electrophysiological responses) while minimizing cortical electrophysiological responses. We first examined the extent to which the FFR decoding accuracies were significantly above chance. Permutations tests showed that decoding accuracies (left panel in Figure 8) were significantly above chance level (indicated by the black triangles) across the four experimental conditions (all *p-value*s = 1.9996×10^{-4}). Next, we examined the effects of visual load and stimulus context on FFR decoding accuracies. There was a significant interaction between visual load and stimulus context ($p = 1.9996 \times 10^{-4}$). Follow-up pairwise comparisons showed that, as displayed in the left panel of Figure 8, in the

predictable context, decoding accuracies were significantly lower for the high visual load condition relative to the low visual load condition (p = 1.9996 × 10^{-4}, uncorrected). However, in the variable context, decoding accuracies were significantly higher for the high load condition than the low load condition (p = 9.998 × 10^{-4}, uncorrected). In the low load condition, decoding accuracies for the predictable context were significantly higher than for the variable context (p = 9.998 × 10^{-4}, uncorrected). But in the high visual load condition, decoding accuracies for the predictable context were significantly lower than for the variable contexts (p = 1.9996 × 10^{-4}, uncorrected).

Feature input of 80-180 Hz

In this analysis, the feature inputs from the FFRs span a narrow frequency band from 80 to 180 Hz that covers F0 range of all the three Mandarin tones (~100 to ~140 Hz). We first examined the extent to which the FFR decoding accuracies were significantly above chance. Permutations tests showed that decoding accuracies (middle panel in Figure 8) were also significantly above chance level (indicated by the blue triangles) across the four experimental conditions (all *p-value*s = 1.9996 × 10^{-4}). Then, we tested the effects of visual load and stimulus context on FFR decoding accuracies. As displayed in the middle panel of Figure 8, the patterns of decoding accuracies were similar to those found for feature inputs of 80-2500 Hz. Statistically, there was a significant interaction between visual load and stimulus context (p = 3.9992 × 10^{-4}). Follow-up pairwise comparisons showed that, in the predictable context, decoding accuracies were significantly lower for the high visual load condition relative to the low visual load condition (p = 1.9996 × 10^{-4}, uncorrected). However, in the variable context, decoding accuracies were marginally significantly higher for the high load condition than the low load condition (p = 0.0688, uncorrected). In the low load condition, decoding

accuracies for the predictable context were significantly higher than for the variable context ($p = 6.399\times 10^{-3}$, uncorrected). But in the high visual load condition, decoding accuracies for the predictable context were significantly lower than for the variable contexts ($p = 2.5995 \times 10^{-3}$, uncorrected).

Feature input of 180-600 Hz

In this analysis, the feature inputs from the FFRs span a narrow frequency band from 180 to 600 Hz that encompasses the second through fourth harmonics (H2-H4) of all the three Mandarin tones. We first examined the extent to which the FFR decoding accuracies were significantly above chance. Permutations tests showed that decoding accuracies (right panel in Figure 8) were barely, but significantly above chance level (indicated by the blue triangles) for three of the four experimental conditions (predictable context + low load: $p = 0.0258$; variable context + low load: $p = 0.0474$; variable context + high load: $p = 0.0258$). Decoding accuracies for the remaining condition (predictable context + high load) were not significantly above chance level ($p = 0.149$). Then, we tested the effects of visual load and stimulus context on FFR decoding accuracies. There was no significant interaction between visual load and stimulus context ($p = 0.139$), or significant main effect of visual load ($p = 0.333$) or stimulus context ($p = 0.335$).

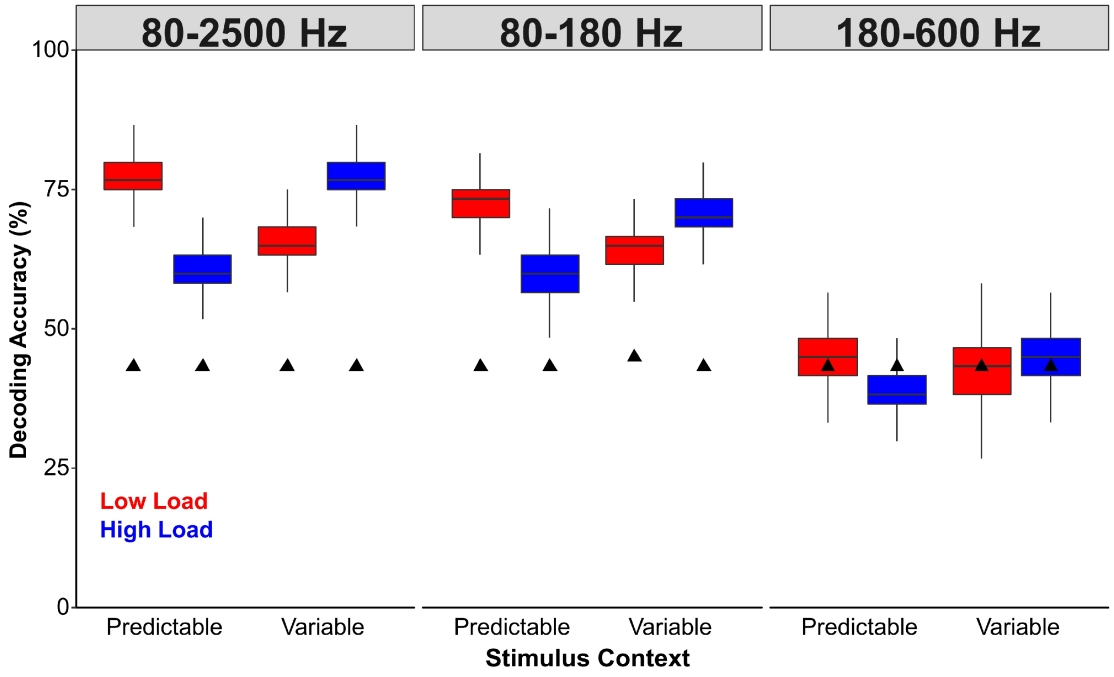

Figure 8. Boxplots of the accuracies to decode information related to the Mandarin tones from the FFRs. The classification analysis used features (500 amplitude values) from the *post*-stimulus region (10 to 100 ms) which covers the frequency range 80 to 2500 Hz (left), 80 to 180 Hz (middle), and 180 to 600 Hz (right), respectively. The auditory stimuli (Mandarin tone T1, T2, or T4) were presented in either predictable or variable contexts, wherein participants performed a visual task of either low (red) or high (blue) visual load. There were 5,000 iterations in each classification analysis, yielding 5,000 decoding accuracies. The boxes and the horizontal line inside shows the quartiles (1st to 3rd quartile) and the median, respectively. The whiskers denote 1.5 times the interquartile range. Outliers, defined as cases with values outside the 1.5 interquartile range, were not displayed here but were included for statistical analysis. The black triangles indicate the 95th percentile of decoding accuracies from permutation tests.

FFRs: Tracking of F0 Contours in the Mandarin Tones

We employed a three-way repeated measures ANOVA to test the effects of visual load and stimulus context on the neural tracking of F0 contours in the Mandarin tones. In this analysis, visual load (high vs. low), stimulus context (predictable vs. variable), and tone (T1, T2, and T4) were included as within-subject factors. We converted stimulus-to-response correlation and peak autocorrelation into Fisher's Z scores to improve the normality of the data and used the converted Z scores for statistical analyses (Wong, Skoe, Russo, Dees, & Kraus, 2007; Xie et al., 2017).

For the stimulus-to-response metric, the main effect of stimulus context was marginally significant [$F(1,19) = 3.97$, $p = 0.061$, $\eta_p^2 = 0.173$], indicating that the mean stimulus-to-response was higher in the predictable context (mean = 0.675, *SD* = 0.249) relative to the variable condition (mean = 0.627, *SD* = 0.234). The interaction between visual load and stimulus context was marginally significant [$F(1,19) = 4.283$, $p = 0.052$, $\eta_p^2 = 0.184$]. Simple effect analysis revealed that, the mean stimulus-to-response was higher in the low load condition (mean = 0.72, *SD* = 0.262) relative to the high load condition (mean = 0.629, *SD* = 0.229) for the predictable context [$t(59) = -2.191$, $p = 0.0324$, uncorrected], but not for the variable context [low load: mean = 0.626, *SD* = 0.257; high load: mean = 0.628, *SD* = 0.212; $t(59) = 0.0369$, $p = 0.971$, uncorrected]. The main effects of visual load or tone, or two-way or three-way interaction between visual load, stimulus context and tone did not reach significance (all *p-value*s > 0.093, η_p^2 ranging from 0.015 to 0.123). For the peak autocorrection metric, none of the main effects, two-way or three-way interaction between visual load, stimulus context and tone were significant (all *p-value*s > 0.109, η_p^2 ranging from 0.013 to 0.12). It is important to note that due to the constraints of using a behavioral task, the number of FFR trials (~95 trials per tone) is extremely low relative to typical studies examining the FFRs (e.g.,

Krishnan et al., 2005, 2004; Xie et al., 2014). Despite this, we see trends in the same direction as the novel machine learning classification metrics.

DISCUSSION

We examined the extent to which visual perceptual load modulates the early sensory representation of speech signals. Our results demonstrate that early sensory representation of linguistically-relevant suprasegmental features (pitch patterns), as indexed by the FFRs, were modulated by the level of perceptual load in the visual task, as well as the context in which the task-irrelevant speech stimuli were presented. When irrelevant speech stimuli were presented in predictable contexts, increasing visual load reduced the decodability of FFRs. However, an opposite pattern was observed when the speech stimuli were presented in variable contexts, such that the decodability of FFRs increased during higher visual load. These findings suggest that focusing attention on a visual task of high perceptual demand influences early auditory encoding, but in a context-dependent manner. The direction of visual attentional influence is highly contingent on the predictability of the incoming auditory stream.

The *supramodal* account of attentional resources posits that visual and auditory processing share central, capacity-limited neural resources, and the depletion of resources by one modality diminishes available resources to the other modality (Broadbent, 1957; Ciaramitaro et al., 2017; Jolicoeur, 1999; Klemen et al., 2009; Macdonald & Lavie, 2011; Molloy et al., 2015; Raveh & Lavie, 2015). Hence, increasing perceptual load on a visual task would lead to reduced availability of neural resources for the processing of task-irrelevant auditory stimuli (Macdonald & Lavie, 2011; Molloy et al., 2015; Raveh & Lavie, 2015). A recent study focusing on auditory cortical processing demonstrated that higher visual load is associated with decreased auditory cortical responses to irrelevant

auditory stimuli (Molloy et al., 2015). Similarly, we demonstrate that early auditory cortical activity, as indexed by the N1 response, is reduced during high visual load, relative to the low visual load (see Figure 5A and 5B). Critically, we demonstrate that the load modulation of auditory processing can be evidenced even at the earliest levels of sensory processing involving stimulus encoding, as indexed by the FFRs (see Figure 8).

Notably, our findings suggest that additional mechanisms are at play in mediating the impact of visual load on early auditory processing. Our results likely reflect a complex interaction between top-down and bottom-up processes in mediating the auditory responsivity to task-irrelevant stimuli. Animal studies suggest that auditory cortical modulation can fine-tune the representation of auditory signals in subcortical nuclei (e.g., Suga, Yan, & Zhang, 1997; Yan & Suga, 1996). Such corticofugal modulatory influence is argued to be important for the selection of regularities in the stimulus stream. Indeed, a prominent role of the auditory cortex is in predictive processing, i.e., making continuous predictions based on prior experience to enhance bottom-up signals. Enhanced fidelity of the FFR in predictive contexts may reflect cortical tuning to enhance the representation of predictable regularities in the incoming stimulus stream. Also, animal models have revealed that subcortical neurons, especially in the inferior colliculus (IC), are highly sensitive to novelty and adapt locally in a stimulus-specific manner. This leads to a decrease in responsivity to repetitive stimuli and heightened responsivity to less predictable stimuli (Anderson & Malmierca, 2013; Malmierca et al., 2009). This form of SSA is predominantly a locally-generated process (Anderson & Malmierca, 2013; Duque & Malmierca, 2015) that gears towards novelty detection.

Based on our results, we posit that there is a constant push-pull between auditory cortical modulation (reflecting predictive processes) and locally-driven processes like

SSA (reflecting novelty detection) in mediating sensory representation, as indexed by the FFR. This argument is supported by a recent animal study demonstrating that IC neurons exhibiting SSA also receive feedback projections from auditory cortex to the IC (Ayala et al., 2015), which provides the infrastructure for dynamic auditory cortical control over local subcortical processes. We posit that the low-load visual task leaves enough resources available for auditory processing, such that the stronger involvement of auditory cortical top-down modulation overrides novelty detection in favor of enhancing predictable elements. Hence, the dominance of auditory cortical control may sharpen the encoding of predictable stimuli at early subcortical levels of processing. When the task involves higher visual perceptual load, all or most of the shared neural resources are consumed, and little or none is left for mediating top-down auditory control. This leads to diminished auditory cortical activity (Molloy et al., 2015), and hence weaker auditory cortical top-down control of the IC (Anderson & Malmierca, 2013; Bajo, Nodal, Moore, & King, 2010; Zhang & Suga, 1997). However, locally-generated processes like SSA are still preserved (Anderson & Malmierca, 2013), and may, in fact, become more dominant. The dominance of local processes may be the norm during sleep, for example, where there is an important benefit for gearing the system towards novelty detection (e.g., waking up to threat). The combined effects of decreased auditory cortical control and preserved SSA at the IC may render the subcortical representation of predictable stimuli less robust but enhance subcortical representation of variable stimuli.

It is assumed that visual processing is prioritized in our tasks (Lavie, 2005; Macdonald & Lavie, 2011; Molloy et al., 2015; Raveh & Lavie, 2015). That is, because participants are instructed to perform the visual search task and ignore the incoming auditory stimuli, attentional resources will first be allocated to visual processing. If there are attentional resources left, these available resources will then be used for auditory

processing. In other words, the visual task would not be affected by the presence of the auditory stimuli. This assumption is in direct contrast to our behavioral finding that the predictability of the auditory stimulus stream influenced performance on the visual task, such that visual targets presented concurrently with unpredictable auditory stimuli (in the variable context) were associated with overall prolonged reaction time. Hence, the assumption that visual processing is prioritized may be violated in the variable context. This may be because variable, unpredictable auditory stimuli may be more distracting and require more resources to process, compared to predictable auditory stimuli (Southwell et al., 2017). This may explain why we observed a different pattern of visual load effect on the FFRs between variable and predictable contexts.

Importantly, we investigated the relevance of features of the FFR for its decodability under different visual load and auditory predictability conditions. Our results indicate that the F0 of the FFRs might be the feature changing as a function of visual load and auditory predictability. First, decoding accuracies of FFRs were well above chance, and influenced by visual load and auditory predictability for FFRs with spectrotemporal information covers the F0 range of all the three tones (~100 to ~140 Hz), but not when the FFRs were filtered above the F0 range (180-600 Hz). The current finding of frequency-specific modulation by visual load is in line with a recent study demonstrating that attention modulated FFRs to stimuli with modulation rates at ~100 Hz, but not to stimuli with modulation rates at above 200 Hz (Holmes, Purcell, Carlyon, Gockel, & Johnsrude, 2018). Further, we directly examined the neural encoding of F0 contours as indexed by the FFRs. Partly consistent with the decoding results, we found a more faithful representation of F0 contours (higher stimulus-to-response correlation) in the low load condition than the high load condition for the predictable context, but not for the variable context. Note that the stimuli used in the current study (T1, T2, and T4) differ

not only in F0 contours, but also in amplitude contours. In our machine learning decoding analysis on the FFR, we used the FFR amplitude values as feature inputs into the classifier, which means that information related to amplitude contours was not excluded. The amplitude contours, in addition to F0, might contribute to the decoding performance.

In a recent MEG study, Coffey et al. (2016) demonstrated a contribution from auditory cortex to the F0 of the FFR close to 100 Hz, in addition to contributions from subcortical nuclei. This raises the possibility that the FFRs recorded in the current study reflect auditory cortical contribution, given that our auditory stimuli have F0s from ~100 to ~140 Hz and substantial energy in these regions in the response. However, the possibility that we are examining cortical representation is unlikely to explain our main findings. First, FFR derived from scalp-recorded EEG (as in the current study) and MEG likely reflect different source contributions (Ahlfors, Han, Belliveau, & Hämäläinen, 2010; Cohen & Cuffin, 1983; Goldenholz et al., 2009). Interpretations regarding sources from MEG cannot be directly applied to EEG. A recent study (Bidelman, 2015a), which used a stimulus with a similar F0 (88-120 Hz) to our study, examined source contributions of the FFR using multichannel scalp-recorded EEG, and indicated that the sources of FFR are consistent with generators in the IC. Second, King et al. (2016) found that FFRs at 85 to 145 Hz (similar to the F0 range of our stimuli) has a latency about 8 to 9 ms, suggesting sources in the rostral brainstem or IC (Møller & Jannetta, 1982). Third, in the present study we found a consistent effect of visual load (reduced N1 amplitude for high vs. low load) on auditory cortical responses across stimulus contexts (Figure 5B). However, the data-driven decoding results do not reflect a simple effect of visual load. We observed an interaction between visual load and stimulus context (see left and middle panels in Figure 8). Based on our results, we suggest that modulating visual load and stimulus context can be utilized as an experimental strategy to evaluate the relative

contribution of multiple top-down and bottom-up processes that influence early speech representation.

The current findings shed light on the controversy over whether cross-modal attention modulates human FFR. The mixed findings may reflect variation in the degree to which listeners disengage their attention from the auditory stimuli (Sörqvist et al., 2012). The current study systematically manipulated the perceptual load of the visual task and provided evidence for the modulation of the FFR by visual attention. Further, task paradigms utilized in previous studies did not require listeners to consistently maintain attention on a simultaneously presented visual task, leaving opportunities for attentional capture by the auditory stimuli. To manipulate visual attention more rigorously, future studies may adapt paradigms similar to the current study, wherein participants are required to consistently focus attention to brief visual stimuli that coincide with the auditory stimuli.

The current study extends recent work demonstrating the feasibility of machine learning approaches to characterize FFR evoked by segmental speech features (vowels) (Sadeghian, Dajani, & Chan, 2015; Yi et al., 2017). Here we utilized a machine learning approach to decode FFRs evoked by speech *suprasegmental* features (linguistically-relevant pitch patterns) (Llanos et al., 2017). Critically, our data showed for the first time that the metric derived from the computational approaches could capture biologically relevant influences (e.g., visual attention) on early sensory encoding. It is important to note that, the decoding approach assumes that FFRs contain information that distinguishes between the auditory stimuli. The decodability of FFRs can be used to index differences in neural representation across stimuli, which can provide novel information about auditory processing beyond activity levels. Our results, combined with Molloy et al. (2015), suggest that visual load not only modulates the overall auditory

activity to task-irrelevant stimuli but also modifies representational differentiation across stimuli.

The current study adopted an experimental design for assessing cortical activity (Molloy et al., 2015), wherein we measured FFRs to speech signals with less than 100 repetitions. With the current design and machine learning analysis approach, we replicated the findings on context effects on FFR (predictable > variable) using traditional approaches (Chandrasekaran et al., 2009; Lau et al., 2016; Parbery-Clark et al., 2011; Slabu et al., 2012). This, to some extent, confirms the applicability of our approach using less FFR trials and decoding analysis. An immediate benefit is the reduction of experimental time and efforts, which will bring extra bring extra convenience to the application of FFR to hard-to-test populations, e.g., young children and older adults. Besides, the number of FFR trials needed for the machine learning approach is within a range similar to the traditional approach to study auditory cortical processing with EEG. That means that the machine learning approach opens up the opportunity to study noninvasively cortical and subcortical auditory processing "truly" simultaneously in human listeners, which would provide insights into the hierarchy of auditory processing along the auditory pathways (Bidelman, 2015b; Shiga et al., 2015).

The FFR is widely considered as a biomarker of auditory function (Johnson, Nicol, & Kraus, 2005; Skoe & Kraus, 2010). A vast majority of studies elicit the FFR by repetitively presenting auditory stimuli to a listener who is also watching a video. The results of the current study pose an important methodological consideration when interpreting results in the FFR literature. Specifically, presenting visual scenes and auditory stimuli simultaneously may affect stimulus encoding as measured by the FFR. Even worse, the extent of visual engagement may vary across sessions and participants,

causing further confounds. Thus, researchers may consider refinements to the practice of FFR collection, including an effective control for the visual components of the task.

In conclusion, our data provide important insights into the mechanisms of multisensory processing. When the brain is overloaded with sensory information from various modalities, the competition for central, capacity-limited perceptual resources among the modalities impacts early encoding of sensory inputs in the task-irrelevant modality. Critically, this influence does not simply result in reduced early sensory processing in the task-irrelevant modality, but rather is dependent on the predictability of the incoming stimulus stream, a possible reflection of the push-pull dynamic between predictive processes and novelty detection within the auditory system.

Chapter 4: Dividing attention to the visual modality impairs the processing of continuous speech

INTRODUCTION

We frequently hear speech in the presence of tasks from sensory modalities other than audition, e.g., listening to the news on the radio while driving. How does dividing attention across modalities affect speech processing? According to the *supramodal* account of attentional resources (Broadbent, 1957; Ciaramitaro, Chow, & Eglington, 2017; Jolicoeur, 1999; Klemen, Büchel, & Rose, 2009; Macdonald & Lavie, 2011; Molloy, Griffiths, Chait, & Lavie, 2015; Raveh & Lavie, 2015), attending to other senses may directly interfere with speech processing. In contrast, based on the *modality-specific* account of attentional resources (Alais et al., 2006; Arrighi et al., 2011; Duncan et al., 1997; Keitel et al., 2013; Parks et al., 2011; Porcu et al., 2014), diverting attention to other modalities may not compromise (auditory) speech processing. Empirically, accumulating behavioral evidence suggest that, engaging attention to visual tasks exerts detrimental effects on many aspects of speech processing (e.g., Hadar et al., 2016; Halin et al., 2015; Hunter & Pisoni, 2017; Mattys et al., 2014; Mattys & Wiget, 2011; Mitterer & Mattys, 2017; Sinnett et al., 2006; Toro, Sinnett, & Soto-Faraco, 2005), supporting the *supramodal* account of attentional resources. To date, however, it is less well understood about the neural mechanisms underlying which dividing attention to visual tasks influences natural, continuous speech processing. Thus, the current study investigates the influence of crossmodal attention on speech processing with continuous speech.

EEG has been one of the primary methods in the investigation of neural processing of speech signals with continuous speech stimuli. A number of EEG studies has focused on the neural representation of temporal envelope in speech signals (e.g., Crosse et al., 2015; Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Fuglsang et al.,

2017; Kong et al., 2014; Lalor & Foxe, 2010; Mirkovic et al., 2015; O'sullivan et al., 2014; Power et al., 2012; Puschmann et al., 2017). The temporal envelope, which reflects fluctuations in amplitude for frequencies between about 2 and 50 Hz, is an important acoustic cue for speech perception (Rosen, 1992). For example, the temporal envelope is robust enough to aid the recognition of vowels, consonants, and sentences when the spectral information is severely reduced (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). To our knowledge, another critical suprasegmental speech feature, the fundamental frequency (F0), has not been examined in EEG responses to continuous speech. F0, a reflection of vocal fold vibration rate between about 50 and 500 Hz, is the primary acoustic correlate of voice pitch (Rosen, 1992). Voice pitch is an important cue for prosodic cues in speech that provides critical linguistic (Rosen, 1992) and paralinguistic (e.g., Rodero, 2011) information. Voice pitch patterns may signal different lexical meaning of a word (Gandour, 1983; Howie, 1976). Voice pitch may also contribute to linguistic processing beyond lexical information. For example, variations in pitch patterns play important roles in resolving ambiguity in syntactic structure (Cutler, Dahan, & Van Donselaar, 1997), and in distinguishing sentences types such as questions and statements (e.g., Gósy & Terken, 1994).

Over the past several years, a number of studies utilizing EEG have begun to examine the neural representation of linguistic features, e.g., phonetic features, in continuous speech (Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Khalighinejad et al., 2017). For example, Khalighinejad et al. (2017) extracted EEG responses time-locked to individual phonemes in continuous speech (called 'phoneme-related potential', PRP). The findings suggest that the PRPs reflect phonological categories (e.g., plosive, fricative, nasal, and vowel) in speech, such that PRPs for the same phonological category

(e.g., vowel) show similar patterns of activity, but differ from that of another phonological category (e.g., plosive, fricative, or nasal).

Here we examined the extent to which dividing attention to a visual task affects neural processing of the suprasegmental (temporal envelope and F0) and segmental (phonetic features) speech features in continuous speech (Figure 9). In a dual-task paradigm, participants performed a primary visuospatial *n*-back task (Jaeggi et al., 2007) with two levels of task demand (0-back, low demand vs. 3-back, high demand) while listening to narrative stories of approximately 60 s long as a secondary task (Figure 4). In a third condition, participants were presented with similar visual-audio stimuli, but were instructed to attend to the auditory stimuli and ignore the visual stimuli (active listening). EEG responses to the stories were recorded across the three task conditions (active listening, visuospatial 0- and 3-back). Based on the *supramodal* account of attentional resources (Broadbent, 1957; Ciaramitaro, Chow, & Eglington, 2017; Jolicoeur, 1999; Klemen, Büchel, & Rose, 2009; Macdonald & Lavie, 2011; Molloy, Griffiths, Chait, & Lavie, 2015; Raveh & Lavie, 2015), we predicted that, diverting attention to the visual task may reduce the behavioral comprehension of the stories and the neural representation of envelope, F0, and phonetic features compared with the active listening condition. Increasing task demand of the visual task (from 0-back to 3-back) may further decrease the behavioral comprehension performance and the neural representation of these three types of speech features. We had no a priori predictions about the extent to which crossmodal attention differentially influences the processing of the three types of speech features.

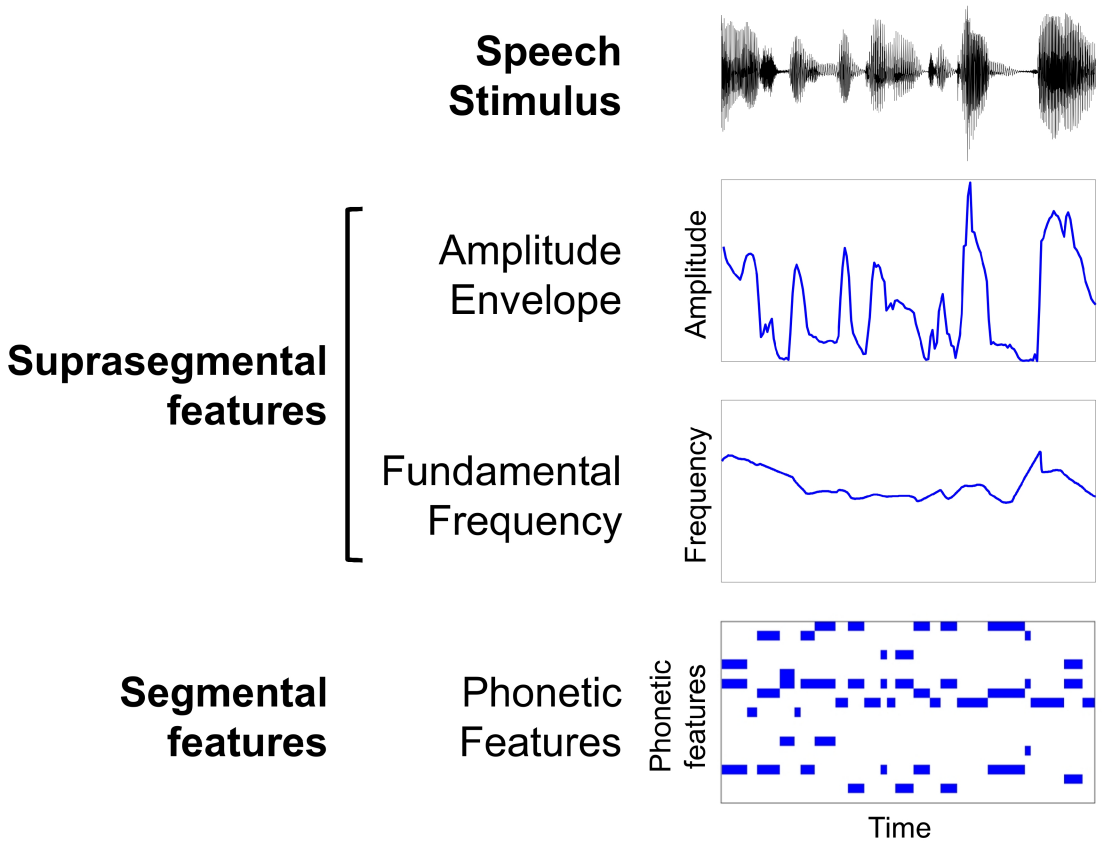

Figure 9. Three types of speech features from the continuous speech stimulus (a segment as an example) examined in study 2: amplitude envelope, fundamental frequency, and phonetic features. The amplitude envelope and fundamental frequency represent suprasegmental features in speech, while the phonetic features represent segmental features in speech. Please refer to the method section for procedures to estimate these speech features.

METHODS

Participants

Eighteen adult native speakers of American English were recruited from the University of Texas at Austin. Data from one participant were excluded due to technical problems. Data from another participant were excluded, because the comprehension accuracy for the stories was lower in the active listening condition (66.67%) than the two visuospatial conditions (0-back: 73.37%; 3-back: 76.67%), indicating that the participant may not have understood the task instructions. Data of the remaining sixteen participants (18 to 23 years old; 11 females, 5 males; 14 right-handed, 2 left-handed) were included in the analysis. Previous evidence has demonstrated that music training influences speech processing (e.g., Bidelman & Alain, 2015; Coffey, Mogilever, & Zatorre, 2017), therefore we recruited participants with either no history or no significant formal music training (<= 4 years of continuous training, not currently practicing). All participants reported no history of psychological or neurological disorders, no use of neuropsychiatric medication, and no prior history of a hearing deficit. Each participant had normal or corrected to normal vision, and underwent audiological screening to ensure that both air and bone-conduction audiometric thresholds were <= 20 dB hearing level (HL) for octave frequencies from 250 to 8,000 Hz, as measured by an Interacoustics Equinox 2.0 PC-Based Audiometer. Each participant provided written, informed consent before the experiment, and received monetary compensation for their participation. The experimental protocol was approved by the Institutional Review Board at the University of Texas at Austin.

Stimuli and Apparatus

The stimuli consisted of visuospatial and continuous speech materials. The visuospatial stimuli were adapted from Jaeggi et al. (2007), and were displayed on a VIEWPixx/EEG scanning LED-backlight LCD monitor (height: 29.1 cm, width: 52.2 cm; display resolution: 1920*1080; refresh rate: 120 Hz), placed ~110 cm from the participants' eyes. As shown in Figure 4, blue squares were presented at one of eight different loci around a white fixation cross in the center of a black screen. Each square appeared for 500 ms, and the interval between consecutive squares was 2500 ms. One trial consisted of 23 blue squares and lasted 69 s. The fixation cross was presented throughout the trial.

The continuous speech stimuli were selected from a classic work of fiction Alice's Adventures in Wonderland (Chapters 1-7, http://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll-5). The audiobook was narrated in English by an adult male speaker of American English and sampled at 22.05 kHz. The recorded chapters were divided into 45 segments (each ~60 s in length). Each segment began where the story ended on the previous segment. Silent periods in each segment that exceeded 500 ms were truncated to 500 ms in duration. The story stimuli were equated for root-mean-square (RMS) amplitude at 70 dB sound pressure level (SPL) and presented diotically via insert earphones (ER-3; Etymotic Research, Elk Grove Village, IL). The SPL levels were measured by presenting sounds via the same insert earphones for experiment to a Brüel & Kjaer artificial ear (type 4152) connected with a Brüel & Kjaer hand-held analyzer (type 2250-L). As displayed in Figure 4, each story segment was presented concurrently with one trial of visuospatial stimuli, wherein the segment began 3 s after the visual trial (i.e., starting at the onset of the second blue square in a trial) and ended earlier than the visual trial.

Task Design and Procedure

Overview

We implemented a within-subject design to compare the neural processing of continuous speech under three conditions: an active listening condition and two levels of visual task demand. An active listening condition served as the baseline condition, where participants were required to attend to the auditory stimuli (story segments) and ignore the visuospatial stimuli. High and low visual task demands were manipulated via 3- and 0-back tasks on the visuospatial stimuli (i.e., blue squares) (Figure 4), respectively. The two visuospatial *n*-back conditions were administered as a dual-task, whereby participants completed the visuospatial *n*-back task as a primary task and were instructed to attend to the auditory stimuli (story segments) as a secondary task. Across all three conditions, two multiple-choice comprehension questions for the story segments were asked at the end of each trial to obtain a behavioral measure of attention modulation on continuous speech processing.

The three task conditions were presented in separate blocks and utilized similar stimulus setup as detailed in section of Stimuli and Apparatus. To maintain the continuity of the storyline, we fixed the order of the story segments but counterbalanced the order of the three conditions across participants. Each condition consisted of 15 trials of visuospatial stimuli paired with 15 unique story segments. The offset differences between the visuospatial stimuli and the auditory stimuli were not significantly different across conditions (one-way ANOVA, $F(2, 42) = 0.00798$, $p = 0.992$). The experiment was controlled with E-Prime 2.0.10 (Schneider et al., 2002).

Active listening task

In the active listening task, participants were instructed to focus on the story segments, and ignore the visuospatial stimuli. Participants were required to keep their eyes open during this task and rest their fixation on a white cross in the middle of the screen. Participants were asked two multiple-choice questions to probe comprehension about the story segments at the end of each trial. Participants had unlimited time to answer the story questions. Feedback about accuracy on the story questions was provided to encourage engagement. This task condition contained 30 story questions in total (15 segments * 2 questions/segment).

Visuospatial 3- and 0-back tasks

For the 3-back condition, participants were instructed to respond whether the current blue square matched the one 3 positions back in the sequence (i.e., appearing at the same location as) (Figure 4A). For the 0-back condition, participants responded whether the current blue square matched a predefined target, which was always the first square in the sequence (Figure 4B). The location of the first square was randomized across trials. For both task conditions, a matched square was considered as a target, and an unmatched one was a non-target. Given the nature of the 3-back task, the targets could appear only starting from the fourth square in the sequence. In other words, on a given trial, targets would be among the last 20 squares in the sequence. The number of targets was set at 6. The remaining 14 squares were assigned as non-targets. To match the 3-back task, the number of targets in the 0-back task was also set at 6, and the targets would not appear until the fourth square in the sequence. Blue square starting from the fourth one was counted as non-targets (14 in total), if it was. For both task conditions, targets and non-targets were determined pseudo-randomly, such that their number was held constant while the location of the targets was randomized across trials.

Participants were instructed to respond only to the targets. Speed and accuracy were emphasized as being equally important in making responses. At the end of each trial, accuracy feedback on the visuospatial task was provided to encourage engagement. The number of actual responses was not significantly different between 3- and 0-back conditions [$t(15) = 0.955$, $p = 0.355$]. Importantly, after receiving feedback on the visual task, similar to the active listening task, participants were asked two multiple-choice comprehension questions for the story segment that was presented concurrently with the visuospatial stimuli. Participants had unlimited time to answer the story questions. No feedback about the performance on the story questions was provided. In total, each task condition consisted of 30 story questions (15 segments * 2 questions/segment).

Critically, to manipulate the priority of the visual and auditory tasks, participants were required to primarily focus on the visual task and then attend to the auditory stimulus with whatever cognitive resources they had left. They were explicitly told that if they did not perform well enough on the visual task, their data could not be used.

Electrophysiological Data Acquisition and Preprocessing

During tasks, participants were seated in a comfortable chair in a dark, acoustically shielded booth. Electroencephalography (EEG) data were recorded from 64 actiCAP active electrodes (Brain Products, Gilching, Munich, Germany) placed in the Easycap recording cap (EasyCap; www.easycap.de). The locations of the electrodes were in accordance with the extended 10-20 system (Oostenveld & Praamstra, 2001). Electrode impedances were kept below 20 kΩ. The EEG data were sampled at 5 kHz, and online referenced to the electrode TP9. A common ground was placed at the Fpz electrode site. The EEG data were amplified and digitized with BrainVision actiCHAmp

amplifier (Brain Products, Gilching, Munich, Germany) linked to BrainVision Pycorder software 1.0.7 (Brain Products, Gilching, Munich, Germany).

The EEG data were preprocessed offline with BrainVision Analyzer 2.0 (Brain Products, Gilching, Munich, Germany). The data were re-referenced to the average of the electrodes TP9 and TP10, and then band-pass filtered using a Butterworth infinite impulse response filter (12 dB/octave, zero phase shift). Next, the data were segmented into epochs that were time-locked to the onset of the auditory story segments. The duration of the epochs matched that of the corresponding segments. To improve computational efficiency, the segmented EEG data were downsampled. Independent component analysis (ICA) was independently applied to EEG data from each of the three task conditions for each participant using the restricted Infomax algorithm (Bell & Sejnowski, 1995). Components related to ocular artifacts were identified and removed via visual inspection of their topographical distribution and activation profile (time course). The remaining components were projected back to EEG electrode space. Finally, the EEG data from each electrode was normalized to ensure zero mean and unit variance for each participant at each task condition. Based on prior work (e.g., Di Liberto & Lalor, 2017; Di Liberto et al., 2015), to assess neural processing of speech envelope and phonetic features, we band-pass filtered the EEG responses from 1 to 15 Hz and downsampled to 128 Hz. A recent electrocorticography (ECoG) study suggests that cortical processing of F0 is reflected in neural activity as high as 150 Hz (Tang, Hamilton, & Chang, 2017). Hence, to assess neural processing of F0 in the current study, we band-pass filtered the EEG responses from 1 to 150 Hz (notched filtered at 60 Hz to minimize line noise) and downsampled to 500 Hz. The number of rejected ICA components were not different across the three task conditions for EEG band-pass filtered

from 1 to 15 Hz [$F(2,30) = 0.918$, $p = 0.41$] or from 1 to 150 Hz [$F(2,30) = 0.591$, $p = 0.56$].

Assessing Neural Processing of Continuous Speech with EEG Responses

Predicting EEG responses from speech features

We use a model-based analysis to assess how well the EEG responses reflect the encoding of the three types of speech features (amplitude envelope, fundamental frequency, and phonetic features) (see Figure 9 for examples). Specifically, as illustrated in Figure 10A, a model is fitted using regularized linear regression to quantify the forward mapping from a speech feature to the EEG responses at each EEG electrode. Then, the fitted model is tested to see how accurately it can predict EEG responses from a novel trial of the same type of speech feature. The Pearson's correlation coefficient between the predicted and the actual EEG responses were calculated to index the accuracy with which we could predict the EEG data from speech features (i.e., prediction accuracy). A leave-one-out cross-validation approach was adopted to assess the EEG prediction performance. A model was trained with data from 14 of the 15 trials and then was used to predict the EEG responses from the remaining trial. This process was repeated until all trials had been predicted against. Single prediction accuracy was then derived by calculating the mean prediction accuracy across all the 15 trials. Higher prediction accuracy was taken as reflective of better neural representation of the corresponding features in the speech stimuli (e.g., Di Liberto & Lalor, 2017; Di Liberto et al., 2015; O'Sullivan, Crosse, Di Liberto, & Lalor, 2017; Puvvada & Simon, 2017).

The model is often referred to as temporal response function (TRF), which can be considered as a filter that quantifies the transformation of a stimulus feature to continuous neural responses by the brain (e.g., Di Liberto & Lalor, 2017; Di Liberto et al., 2015;

O'Sullivan et al., 2017). Nonzero weights of parameters can be interpreted as that there is cortical activity is related to the encoding of stimulus (Haufe et al., 2014). The TRF analysis was applied to data at each task condition for individual participants. The TRF analysis was implemented using the multivariate temporal response function (mTRF) MATLAB (The MathWorks, Natick, MA) toolbox (Crosse, Di Liberto, Bednar, & Lalor, 2016).

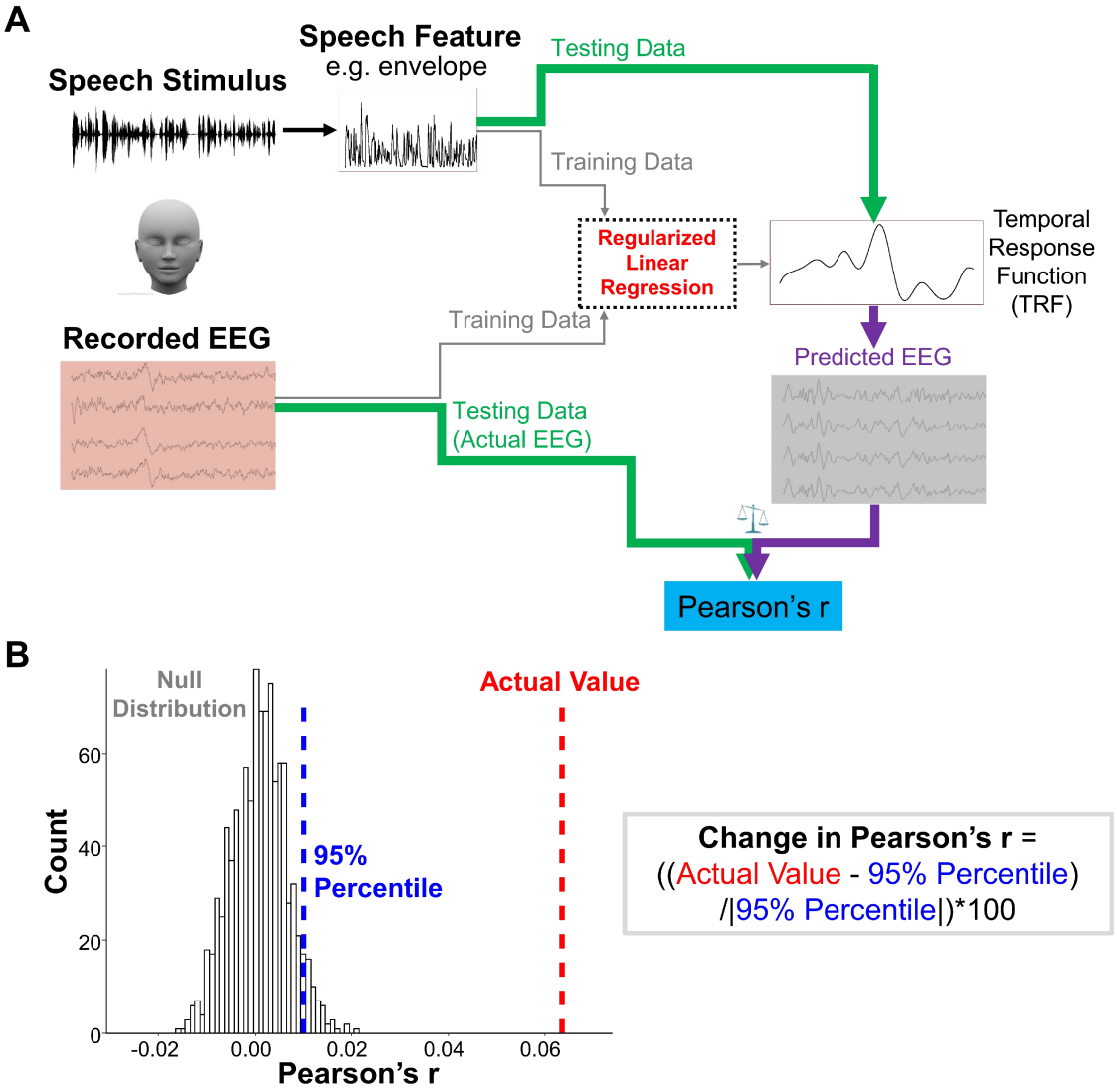

Figure 10. (A) Schematic of the procedures for the temporal response function (TRF) analysis. To determine the significance level of the prediction accuracy (i.e., Pearson's r), we shuffled the stimulus representations (speech features) and conducted a pseudo-TRF analysis on the shuffled stimulus and the actual EEG responses (not shuffled). This shuffling and pseudo-TRF analysis was iterated 1,000 times, and a null distribution of the prediction accuracy was obtained. (B) An example of results from one experimental condition for one participant when amplitude envelope was used as the speech feature. The black histogram represents the distribution of chance-level Pearson's r. The red dashed line represents the actual Pearson's r. The blue dashed line represents the 95th percentile in the chance-level distribution. If the actual Pearson's r is higher than the 95th percentile in the chance-level distribution, we take that as indicating the actual value is significantly above chance. We calculated *Change in Pearson's r* as the measure.

Estimating speech features from the continuous speech stimuli

We estimated TRF based on three distinct types of features from the continuous speech stimuli (Figure 9):

1. Amplitude envelope: This was extracted using a Hilbert transform, via the *hilbert* function in MATLAB (The MathWorks, Natick, MA). The Hilbert transform returns a complex analytic signal that is the sum of the original signal and its Hilbert transform. The speech envelopes were calculated as the absolute value of the analytic signal and were then downsampled to the same sampling rate as the EEG data (i.e., 128 Hz) using the *resample* function in MATLAB (The MathWorks, Natick, MA).
2. Fundamental frequency (F0): This was extracted using *ProsodyPro* version 5.7.7 (Xu, 2013) implemented in *PRAAT* (Boersma & Weenink, 2013). The sampling rate for F0 extraction was set to 500 Hz to match that of the EEG data. F0 values falling in the silence and pause periods, as computed by the *Prosodylab-Aligner* (Gorman, Howell, & Wagner, 2011), were adjusted to be 0.
3. Phonetic features: This type of speech feature was computed using the *Prosodylab-Aligner* (Gorman et al., 2011). Specifically, each word is partitioned into phonemes from the American English International Phonetic Alphabet (IPA). The *Prosodylab-Aligner* then estimated the time-points corresponding to the onset and offset of each phoneme using forced-alignment (Gorman et al., 2011). Each phoneme was projected to a space of 19 phonetic features that describe the manner of articulation, the voicing of a consonant, the backness of a vowel, and the place of articulation (Di Liberto & Lalor, 2017; Di Liberto et al., 2015; Mesgarani, Cheung, Johnson, & Chang, 2014).

Then, this information was converted into a multivariate time-series that constitutes 19 indicator variables (one for each phonetic feature). Each indicator variable is a binary array, such that value '1' is assigned to the time-points when the indicated phonetic feature occurred and value '0' for the time-points when the indicated phonetic feature was not present. Since each phoneme is described by a combination of distinct phonetic features, the indicator variables are not mutually exclusive that multiple indicators may be active (having a value of 1) for one time-point.

Parameters for TRF analysis

Three parameters are critical to the TRF analysis: 1) the number of EEG electrodes; 2) the time lags between stimulus envelope and EEG responses; 3) the regularization parameter λ (Crosse et al., 2016). In terms of the number of EEG electrodes, two options were adopted: a) all the 62 electrodes; b) a set of 10 frontotemporal electrodes (5 on the left side of the scalp, and their symmetrical counterparts on the right) with the highest prediction correlations based on data from another unpublished dataset in our lab. The EEG prediction correlations were averaged across these two choices of electrodes, respectively. The specified time lag was restricted to lags from 0 to 250 ms. This range of time lags has been shown to robustly capture the relationship between various types of speech representations (e.g., envelope and phonetic features) and EEG data (Di Liberto & Lalor, 2017; Di Liberto et al., 2015). Finally, the purpose of including the regularization parameter λ was to prevent overfitting of the model (Crosse et al., 2016). We conducted a parameter search (over the range 10^{-15}, 10^{-14}, 10^{-13}, …,10^{15}) for the λ value that optimized the EEG prediction performance for each

task condition in individual participants. The procedures for tuning the λ value has been extensively described in Crosse et al. (2016).

Testing the significance of prediction accuracy

To determine the significance level of the prediction accuracy, we shuffled the stimulus representations (speech features) and conducted a pseudo-TRF analysis on the shuffled stimulus and the actual EEG responses (not shuffled). This shuffling and pseudo-TRF analysis was iterated 1,000 times, and a null distribution of the prediction accuracy was obtained (see Figure 10B for an example). We then tested the actual prediction accuracy against this null distribution, and estimated the *p* value using the formula: $p = (a+1)/(n+1)$ (Phipson & Smyth, 2010), where a is the number of prediction accuracies from the null distribution that exceeds the actual prediction accuracy, and n is the total number of prediction accuracies from the null distribution (i.e., 1,000). We performed this significance testing for each task condition in individual participants.

Decoding phonetic features from the EEG responses

Recent work suggests that EEG responses time-locked to phonemes (phoneme-related potentials, PRPs) from continuous speech stimuli reflect the encoding of phonetic features (Khalighinejad et al., 2017). The PRPs show a distinction across phonological categories of plosive, fricative, nasal, and vowel (also see an ECoG study Mesgarani et al., 2014). In light of these findings, we aimed to further evaluate the neural processing of phonetic features by decoding phonological categories (plosive, fricative, nasal, and vowel) from PRPs and examined the extent to which diverting attention to the visual task influence the decoding performance. The procedures for the decoding analysis are illustrated in Figure 11.

Extraction of PRPs

As illustrated in Figure 11A, to obtain a time-locked EEG response to each phoneme, the EEG data was segmented and aligned to phoneme onset with a time window of 0 to 600 ms. The phonemes and onset information were computed using the procedures described in the section of *Estimating speech features from the continuous speech stimuli*. PRP was calculated by averaging all the instances of the segmented EEG responses to each phoneme. Examples of the PRPs are displayed in Figure 11B.

Procedures for decoding phonological categories from PRPs

We used a supervised machine learning algorism (linear SVM; Cristianini & Shawe-Taylor, 2000), implemented using the *fitcecoc* function in MATLAB (The MathWorks, Natick, MA). The linear SVM uses a "one-against-one" approach (Knerr et al., 1990). Specifically, as there were four phonological categories (plosive, fricative, nasal, and vowel) in our experiment, the linear SVM constructed 6 classifiers to test the PRP data from all the pairwise combinations of the four phonological categories. The label of the phonological category with the highest probability was taken as the classified label. Default values of all the parameters were used.

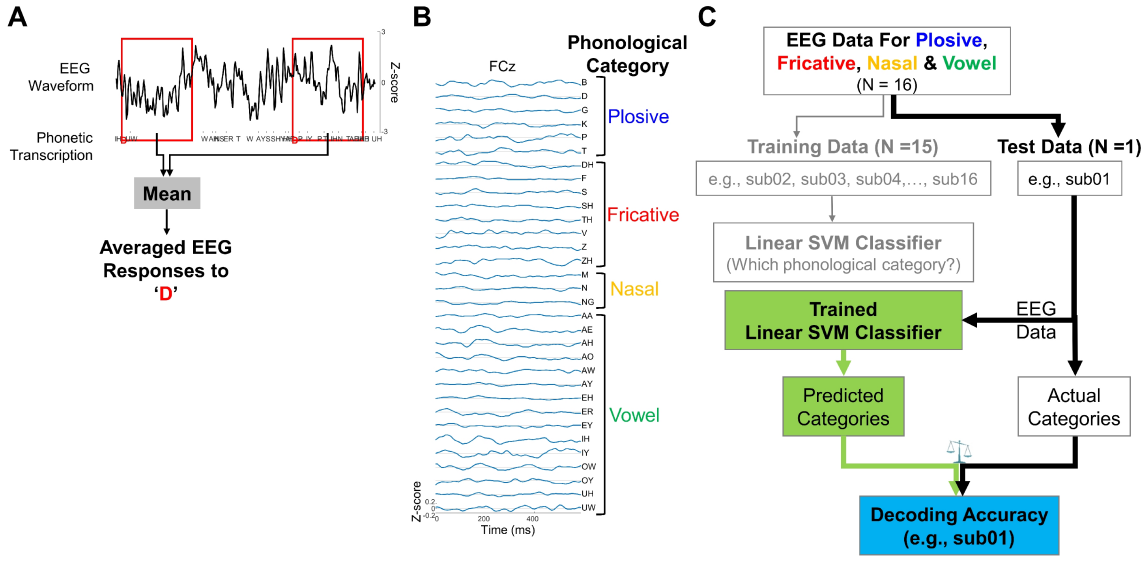

Figure 11. Schematic of procedures for the analysis to decode phonological categories (plosive, fricative, nasal, and vowel) from EEG responses time-locked to phonemes (phoneme-related potentials, PRPs) in continuous speech stimuli. (A) Extraction of PRPs aligned to phoneme onset with a time window of 0 to 600 ms. (B) Examples of the PRPs at electrode FCz from one experimental condition for one participant. The corresponding phonemes are grouped into phonological categories plosive, fricative, nasal, and vowel. (C) Procedures for decoding phonological categories from PRPs.

To evaluate the decoding performance for each participant, as illustrated in Figure 11C, we trained the classifier with the PRP data from 15 of the 16 participants, and the trained classifier was used to predict the phonological category of the PRP data the remaining participant. To handle the issue of uneven number of phonemes across phonological categories (6 for plosive, 8 for fricative, 3 for nasal, and 15 for vowel), we balanced the number for each phonological category using the *cosmo_balance_partitions* function from the *CoSMoMVPA* toolbox in MATLAB (The MathWorks, Natick, MA). This function generated multiple pairs of training and testing dataset, with RPR samples from each phonological category occurs at least once across the training dataset. The

decoding accuracy (i.e., the percentage that the trained classifier correctly predicts the phonological category labels of the PRP data in the testing dataset) was calculated by averaging the accuracies across all the cross-validations. Decoding accuracy was estimated for each of the 62 electrodes. Based on Khalighinejad et al. (2017), we selected a set of 7 frontocentral electrodes where the PRP data showed highest distinctions of phonological categories of plosive, fricative, nasal, and vowel. The decoding accuracies were then averaged across the 7 selected electrodes. We estimated the decoding accuracy for each of the three task conditions respectively.

Testing the significance of decoding accuracy

To determine the significance level of the decoding accuracy, we randomly assigned the labels of the phonological category in the training data, and trained the classifier with the shuffled PRP data, and then predict the phonological category labels of testing data. This label shuffling and decoding analysis was iterated 10 times for each participant (n = 16) and each electrode (n = 62), and a null distribution of the decoding accuracy (n = 10*16*62 = 9,920) was obtained. We then tested the actual decoding accuracy against this null distribution, and estimated the *p* value using the formula: $p = (a+1)/(n+1)$ (Phipson & Smyth, 2010), where a is the number of decoding accuracies from the null distribution that exceeds the actual decoding accuracy, and n is the total number of decoding accuracies from the null distribution (i.e., 9,920). We performed this significance testing for each task condition in individual participants.

Statistical Analysis

We calculated three pieces of data from the current study. First, behavioral performance on the visuospatial *n*-back tasks was assessed by two metrics: accuracy,

which was calculated as the difference in hit rates (i.e., correctly responding to a target) and false alarm rates (i.e., identifying a non-target as being a target); mean reaction times (RTs) for hits only (e.g., Jaeggi et al., 2007; Snodgrass & Corwin, 1988). Second, behavioral performance on the continuous speech stimuli (i.e., story segments) was assessed as the proportion of correctly answered story questions. Third, the cortical (neural) processing of the continuous speech stimuli was quantified in terms of the EEG prediction accuracy (i.e., Pearson's correlation coefficient between the estimated and the actual EEG responses) and the decoding accuracy of phonological categories from PRPs.

We employed similar analysis approaches to examine the effect of visual task demand for the three datasets described above. Specifically, linear mixed-effects model, implemented via the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) in R version 3.4.0 (Team, 2017), was used to fit the data. In the model, visual task demand (treated as a categorical variable) was included as the fixed effects, and by-subject intercept was included as random effects to account for baseline performance difference. For the analysis of the behavioral performance on the visuospatial n-back tasks, there were two levels (0-back and 3-back) for the variable of visual task demand. For the analyses on the behavioral as well as neural performance on the continuous speech stimuli, the variable of visual task demand consisted of three levels (active listening, 0-back, and 3-back). We tested the main effect of visual task demand by comparing the base model (which only included the random-effects structure) to the same model but with the addition of visual task demand. Model comparisons were achieved using the likelihood ratio (Baayen, Davidson, & Bates, 2008). Post hoc analysis for significant main effect, if necessary, was carried out by Tukey's tests using the *glht* function of the *multcomp* package (Hothorn, Bretz, & Westfall, 2008). Multiple comparisons were

corrected using the Benjamini-Hochberg false discovery rate method (Benjamini & Hochberg, 1995).

RESULTS

Behavioral Performance on the Visuospatial n-back Tasks

Figure 12A and 12B display the behavioral performance on the visuospatial 0- and 3-back tasks for individual participants (n = 16). On average, participants responded to the targets with an accuracy of 63.31% (*SD* = 21.85) in the 3-back condition and 99.54% (*SD* = 0.82) in the 0-back condition, and with RT of 785.24 ms (*SD* = 233.41) in the 3-back condition and 453.11 ms (*SD* = 67.54) in the 0-back condition. Statistically, there was a significant main effect of visual task demand for both accuracy [χ^2 (1) = 28.861, $p = 7.777 \times 10^{-8}$] and RT measures [χ^2 (1) = 24.179, $p = 8.779 \times 10^{-7}$]. These behavioral results confirmed that the manipulation of visual task demand was successful, such that high (relative to low) visual task demand was associated with lower accuracy and slower RT.

Behavioral Performance on the Continuous Speech Stimuli

Figure 12C illustrates the behavioral performance on the continuous speech stimuli (i.e., story segments) for individual participants (n = 16). On average, participants correctly answered 88.96% (*SD* = 5.93) of story questions in the active listening condition, 84.58% (*SD* = 11.86) in the visuospatial 0-back condition, and 65.63% (*SD* = 12.75) in the visuospatial 3-back condition. Statistically, there was a significant main effect of visual task demand [$\chi^2(2) = 39.614$, $p = 2.5 \times 10^{-9}$]. Post hoc analysis revealed that, the accuracy on the story questions was significantly lower in the 3-back condition relative to the other two conditions (3-back vs. 0-back, $\beta = -0.1896$, $SE = 0.029$, $Z = -6.538$, $p = 9.36 \times 10^{-11}$; 3-back vs. active listening, $\beta = -0.233$, $SE = 0.029$, $Z = -8.047$, p

$= 2.66 \times 10^{-15}$). The accuracy on the story questions did not significantly differ between active listening and 0-back conditions ($\beta = -0.0438$, $SE = 0.029$, $Z = -1.509$, $p = 0.131$).

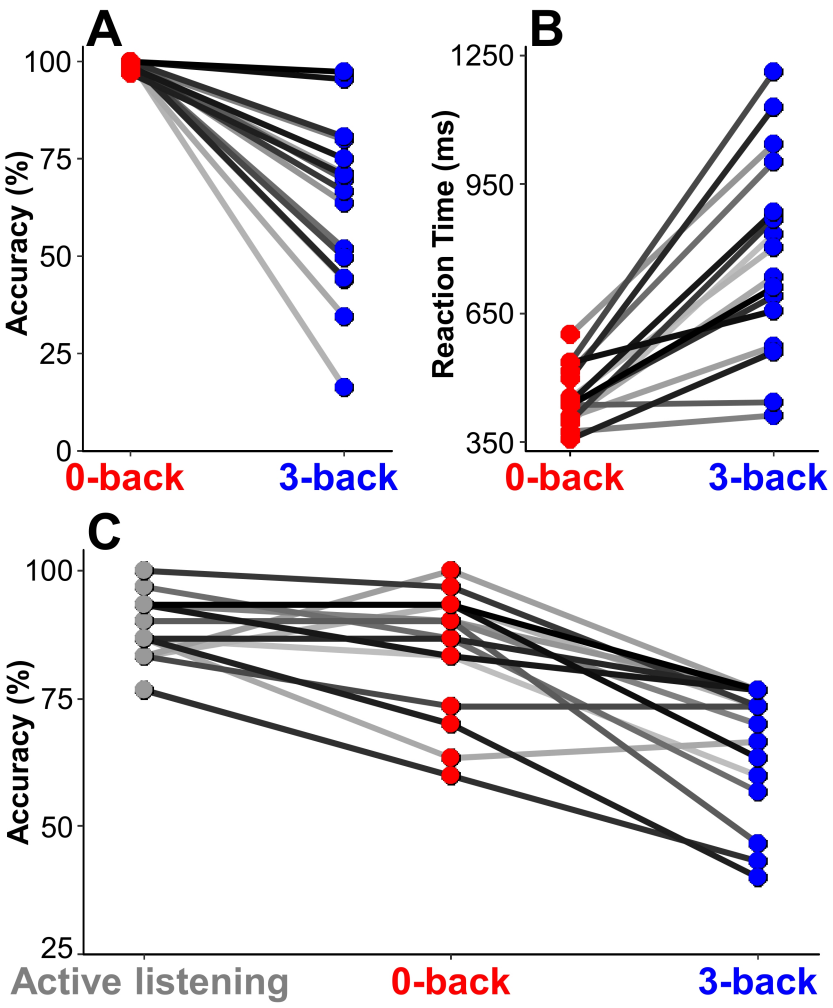

Figure 12. Behavioral performance on the visuospatial 0- and 3-back tasks, and continuous speech stimuli. (A) Accuracy on the visuospatial tasks, which was the difference in hit rates (i.e., correctly responding to a target) and false alarm rates (i.e., identifying a non-target as being a target). (B) Reaction time on the visuospatial tasks for hits only. (C) Accuracy on the questions for continuous speech stimuli, i.e., the proportion of correctly answered story questions. Individual lines denote individual participants (n = 16).

Neural Processing of Continuous Speech Stimuli

Amplitude envelope

The results for amplitude envelope are displayed in the top row of Figure 13. Qualitatively, based on the topographic distributions (the second, third, and fourth plots) of the EEG prediction accuracies (i.e., Pearson's r), the frontotemporal electrodes show the highest prediction accuracies across the three conditions. First, we examined EEG prediction accuracies that were averaged across all electrodes. The prediction accuracies for all conditions and all participants were significantly above chance level (all *p-value*s < 0.05). The fifth plot in the top row of Figure 13 displays the grand-average (n = 16) prediction accuracies for the three conditions. Statistically, there was a significant main effect of visual task demand [$\chi^2(2) = 6.386$, $p = 0.0411$]. Post hoc analysis revealed that, prediction accuracy was (marginally) significantly lower in the visuospatial 3-back and 0-back conditions compared to the active listening condition (3-back vs. active listening, $\beta = -0.00451$, $SE = 0.00238$, $Z = -1.896$, $p = 0.087$; 0-back vs. active listening, $\beta = -0.00584$, $SE = 0.00238$, $Z = -2.456$, $p = 0.0422$). Prediction accuracy was not significantly different between the 3- and 0-back conditions ($\beta = 0.00133$, $SE = 0.00238$, $Z = 0.56$, $p = 0.575$).

Second, as shown in the sixth plot in the top row of Figure 13, we found a similar pattern of results when focusing on selected electrodes (n = 10). The prediction accuracies for all conditions and all participants were significantly above chance level (all *p-value*s < 0.05). There was a significant main effect of visual task demand [$\chi^2(2) = 8.754$, $p = 0.0126$]. Post hoc analysis revealed that, prediction accuracy was significantly lower in the visuospatial 3-back and 0-back conditions compared to the active listening condition (3-back vs. active listening, $\beta = -0.00796$, $SE = 0.00295$, $Z = -2.7$, $p = 0.0132$; 0-back vs. active listening, $\beta = -0.00772$, $SE = 0.00295$, $Z = -2.619$, $p = 0.0132$).

Prediction accuracy was not significantly different between the 0- and 3-back conditions (β = -0.000239, SE = 0.00295, Z = -0.081, p = 0.935).

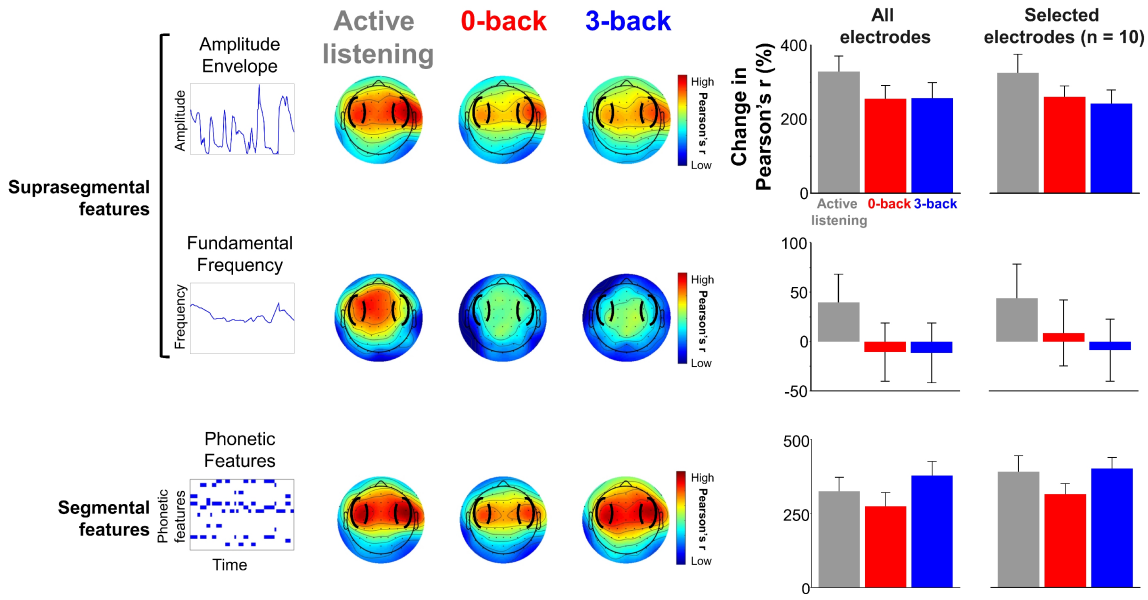

Figure 13. Results for the neural processing of amplitude envelope (top row), fundamental frequency (middle row), and phonetic features (bottom row) in continuous speech stimuli. The second to fourth plots in each row show the topographic distributions of the EEG prediction accuracies (i.e. Pearson's r) across the three task conditions (active listening, 0- and 3-back). Black dashed lines enclose the 10 frontotemporal electrodes selected for analysis. The fifth and sixth plots in each row show the grand-average (n = 16) of change in EEG prediction accuracy (Pearson's r) averaged across all electrodes and selected electrodes for the three task conditions. As shown in Figure 9B, the change in EEG prediction accuracy was calculated as follows: Obtaining the difference between the actual EEG prediction accuracy and the 95th percentile of prediction accuracy in the null distribution, dividing the difference by the absolute value of the 95th percentile of the null distribution of prediction accuracy, and then multiplying the quotients by 100. For statistical analysis, we used the original EEG prediction accuracy. But note that we obtained similar patterns of results when using the metric of change in the EEG prediction accuracy. Error bars denote 95% confidence interval.

Fundamental frequency (F0)

The results for fundamental frequency are displayed in the middle row of Figure 13. Qualitatively, based on the topographic distributions (the second, third, and fourth plots) of the EEG prediction accuracies (i.e., Pearson's r), the frontocentral electrodes show the highest prediction accuracies across the three conditions. First, we examined EEG prediction accuracies that were averaged across all electrodes. The prediction accuracies for 13 participants in the active listening condition, 4 participants in the visuospatial 0-back condition, and 8 participants in the visuospatial 3-back condition were significantly above chance level (ps < 0.05) or marginally significantly above chancel level ($0.05 \leq p$s < 0.06). The fifth plot in the middle row of Figure 13 displays the grand-average (n = 16) prediction accuracies for the three conditions. There was a significant main effect of visual task demand [$\chi^2(2) = 9.785$, $p = 0.0075$]. Post hoc analysis revealed that, prediction accuracy was significantly lower in the visuospatial 3-back and 0-back conditions compared to the active listening condition (3-back vs. active listening, $\beta = -0.00466$, $SE = 0.00165$, $Z = -2.824$, $p = 0.0105$; 0-back vs. active listening, $\beta = -0.00445$, $SE = 0.00165$, $Z = -2.696$, $p = 0.0105$). Prediction accuracy was not significantly different between the 0- and 3-back conditions ($\beta = -0.000212$, $SE = 0.00165$, $Z = -0.128$, $p = 0.898$).

Second, as shown in the sixth plot in the middle row of Figure 13, we found a similar pattern of results when focusing on selected electrodes. The prediction accuracies for 11 participants in the active listening condition, 7 participants in the visuospatial 0-back condition, and 10 participants in the visuospatial 3-back condition were significantly above chance level (ps < 0.05) or marginally significantly above chancel level ($0.05 \leq p$s < 0.06). There was a significant main effect of visual task demand [$\chi^2(2) = 7.227$, $p = 0.027$]. Post hoc analysis revealed that, prediction accuracy was significantly or

marginally significantly lower in the visuospatial 3-back and 0-back conditions compared to the active listening condition (3-back vs. active listening, β = -0.00611, *SE* = 0.00237, *Z* = -2.58, *p* = 0.0296; 0-back vs. active listening, β = -0.00471, *SE* = 0.00237, *Z* = -1.991, *p* = 0.0698). Prediction accuracy was not significantly different between the 0- and 3-back conditions (β = -0.0014, *SE* = 0.00237, *Z* = -0.59, *p* = 0.555).

Phonetic features

The results for phonetic features are displayed in the bottom row of Figure 13. Qualitatively, based on the topographic distributions (the second, third, and fourth plots) of the EEG prediction accuracies (i.e., Pearson's r), the frontotemporal electrodes show the highest prediction accuracies across the three conditions. First, we examined EEG prediction accuracies that were averaged across all electrodes. The prediction accuracies for all conditions and all participants were significantly above chance level (all *p-value*s < 0.05). The fifth plot in the bottom row of Figure 13 displays the grand-average (n = 16) prediction accuracies (Pearson's r) for the three conditions. There was a significant main effect of visual task demand [$\chi^2(2) = 8.6155$, *p* = 0.0135]. Post hoc analysis revealed that, prediction accuracy was significantly or marginally significantly lower in the visuospatial 0-back condition compared to the active listening and 3-back conditions (0-back vs. active listening, β = -0.00596, *SE* = 0.00288, *Z* = -2.07, *p* = 0.0577; 0-back vs. 3-back, β = 0.00855, *SE* = 0.00288, *Z* = 2.968, *p* = 0.00898). Prediction accuracy was not significantly different between the active listening and 3-back conditions (β = 0.00259, *SE* = 0.00288, *Z* = 0.898, *p* = 0.369).

Second, as shown in the sixth plot in the bottom row of Figure 13, we found a similar pattern of results when focusing on selected electrodes. The prediction accuracies for all conditions and all participants were significantly above chance level (all *p-value*s

< 0.01). There was a significant main effect of visual task demand [$\chi^2(2) = 8.281$, $p = 0.0159$]. Post hoc analysis revealed that, prediction accuracy was significantly lower in the visuospatial 0-back condition compared to the active listening and 3-back conditions (0-back vs. active listening, $\beta = -0.00943$, $SE = 0.00347$, $Z = -2.713$, $p = 0.02$; 0-back vs. 3-back, $\beta = 0.0084$, $SE = 0.00347$, $Z = 2.417$, $p = 0.0235$). Prediction accuracy was not significantly different between the active listening and 3-back conditions ($\beta = -0.00103$, $SE = 0.00347$, $Z = -0.296$, $p = 0.766$).

Further, we decoded phonological categories (plosive, fricative, nasal, and vowel) from EEG responses time-locked to the phonemes (PRPs) and examined the effect of visual task demand on the decoding performance. The top panel in Figure 14 displays the topographic distributions of the accuracies to decode phonological categories from PRPs. Qualitatively, the frontocentral electrodes show the highest prediction accuracies across the three conditions. We calculated the decoding performance by averaging across 7 frontocentral electrodes that show highest temporal separation across phonological categories from Khalighinejad et al. (2017). Across the three conditions, decoding accuracies from at least 9 (out of 16) participants (active listening: 9/16; 0-back: 10/16; 3-back: 13/16) were significantly above chance level (ps < 0.05) or marginally significantly above chancel level ($0.05 \leq p$s < 0.06). The bottom panel in Figure 14 displays individual and the distribution of decoding accuracies across the three conditions. Statistically, there was no significant main effect of visual task demand [$\chi^2(2) = 1.054$, $p = 0.59$].

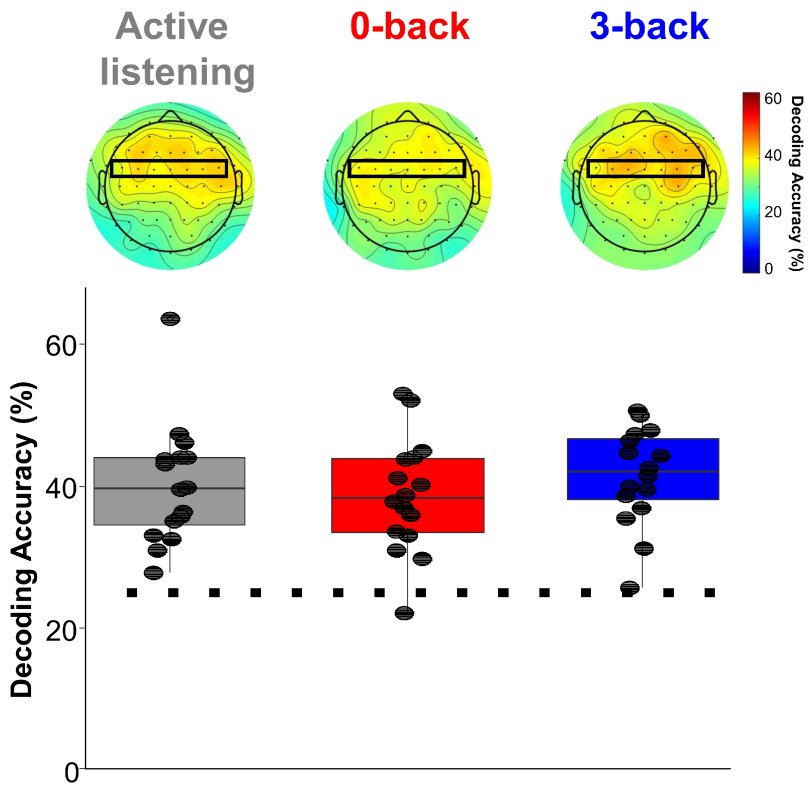

Figure 14. Results for the neural processing of phonetic features in continuous speech stimuli based on the analysis to decode phonological categories (plosive, fricative, nasal, and vowel) from EEG responses to the phonemes (phoneme-related potentials, PRPs). Top: Topographic distribution of the decoding accuracies. Black squares enclose the 7 frontocentral electrodes selected for analysis based on Khalighinejad et al. (2017). Bottom: Boxplots of decoding accuracies. The boxes and the horizontal line inside shows the quartiles (1st to 3rd quartile) and the median, respectively. The whiskers denote 1.5 times the interquartile range. Outliers, defined as cases with values outside the 1.5 interquartile ranges, were not displayed here but were included for statistical analysis. The dots denote individual participants. The black dashed line indicates theoretical chance level (25%).

Relationship between Behavioral and Neural Measures on the Processing of Continuous Speech Stimuli

We investigated whether neural measures on the processing of continuous speech stimuli was related to behavioral performance on the continuous speech stimuli. To

correct for multiple tests, we set the α value at 0.05/21 = 0.0023801, wherein 21 corresponded to the number of correlation tests. We found that, across measures from amplitude envelope, fundamental frequency, and phonetic features (for both average across all electrodes and selected electrodes), change in prediction accuracy was not significantly correlated with behavioral accuracy change on the continuous speech stimuli for the comparison between active listening and 3-back condition [calculated as: (active listening - 3-back)/active listening] (Spearman's rank correlation *rho* ranging from -0.413 to 0.101, *p-value*s ranging from 0.111 to 0.944), for the comparison between active listening and 0-back condition [calculated as: (active listening - 0-back)/active listening] (Spearman's rank correlation *rho*s ranging from -0.493 to 0.0269, *p-value*s ranging from 0.0525 to 0.921), or for the comparison between 0-back and 3-back condition [calculated as: (0-back - 3-back)/ 0-back] (Spearman's rank correlation *rho*s ranging from -0.552 to 0.347, *p-value*s ranging from 0.0263 to 0.345). Further, the accuracy to decode phonological categories from the EEG data was not significantly correlated with behavioral accuracy change on the continuous speech stimuli for any of the comparison between conditions (Spearman's rank correlation *rho*s ranging from -0.165 to 0.107, *p-value*s ranging from 0.541 to 0.846).

DISCUSSION

The present study examined the influence of dividing attention to a visual task on the processing of continuous speech. The visual task used identical visuospatial stimuli and successfully achieved the manipulation of high (3-back) and low (0-back) task demands, as demonstrated by the finding that high, relative to low task demand, resulted in lower accuracy and slower reaction time in responding to the visual targets. Importantly, at the behavioral level, dividing attention to the visual task led to lower

comprehension accuracy on the continuous speech stimuli compared to active listening conditions, but only when the task demand of the visual task was high (3-back but not 0-back). At the neural level, the representation of speech envelope and fundamental frequency (F0) in continuous speech, indexed by the EEG prediction accuracy, was reduced when diverting attention to the visual task. Intriguingly, unlike the behavioral performance, this reduction was observed independently of the task demand of the visual task. In contrast, the representation of phonetic features in continuous speech, indexed by the EEG prediction accuracy and the accuracy for decoding phonetic categories from EEG, was generally unaffected by the manipulation of attention. Taken together, in general agreement with our predictions, these results demonstrated that dividing attention to a visual task impairs the processing of continuous speech.

To our knowledge, our study represents one of the first attempts to investigate crossmodal attention effects on speech processing with more ecologically valid, continuous speech. In a behavioral study, Halin et al. (2015) manipulated crossmodal attention by asking listeners to engage in a visual letter *n*-back task (1-back vs. 2-back) while ignoring a background story. They found that, in a follow-up surprise memory test, listeners remember less of the stories in the 2-back condition than the 1-back condition. Our study extends those findings by showing that, even in a dual-task situation, wherein listeners were asked to attend to the stories as a secondary task, the behavioral performance on the continuous speech was reduced with increasing demand from the visual task (3-back < 0-back). At the neural level, Kong et al. (2014) compared neural responses to speech envelope in continuous speech between an active listening condition and a condition that participants watched a silent movie and ignored the speech stimuli, and found similar envelope-related neural responses between the two conditions. These results were in direct contrast to our findings that shifting attention to a visual task

compromises the neural representation of speech envelopes. This discrepancy may lie in the manipulation of visual attention between our study and Kong et al. (2014). The visual task in our study required active responses from the participants and thus may have relatively more control of visual attention compared to the movie watching task in Kong et al. (2014). Relatedly, the visual task in our study, particularly the visuospatial 3-back task, may be a relatively more difficult task than the movie watching task in Kong et al. (2014).

To the best of our knowledge, few studies have examined the influence of crossmodal attention on the neural processing of F0 with continuous speech stimuli. There is evidence to suggest that the processing of F0 may be pre-attentive. For example, mismatch negativity (MMN), an ERP component indexing early, preattentive stages of cortical processing, was elicited to changes in linguistically-relevant pitch patterns (e.g., Chandrasekaran, Krishnan, & Gandour, 2007; Xi, Zhang, Shu, Zhang, & Li, 2010). These findings are consistent with an early behavioral report by Cherry (1953) that listeners can accurately identify the gender of the talker in the unattended ear, where the primary acoustic cue is F0. However, it has also been shown that MMN to pitch changes is susceptible to attentional manipulations, such as its amplitude is attenuated when participants are engaged in a demanding task from modalities other than audition (e.g., Sussman, 2013; Yucel, Petty, McCarthy, & Belger, 2005). To date, only one recent study used continuous speech stimuli and examined the influence of auditory selective attention on the auditory subcortical responses to F0, and found that the subcortical activity to F0 was reduced for ignored speech (Forte, Etard, & Reichenbach, 2017). Our results extended these findings by showing that, taking attention away from the auditory modality may also be detrimental to the processing of F0 in continuous speech stimuli.

Regarding the processing of phonetic features, there is behavioral and neuroimaging evidence to suggest that its processing is substantially reduced for unattended speech signals (e.g., Mattys et al., 2014; Mattys & Wiget, 2011; Mitterer & Mattys, 2017). In a series of behavior studies, Mattys and colleagues asked participants to perform a concurrent visual task (e.g., searching target red square among distracting colored shapes) while responding to phoneme perception tasks (e.g., discriminating phonemes with different voice-onset time). They found that performance on the phoneme tasks was impaired by the current visual tasks (e.g., Mattys et al., 2014; Mattys & Wiget, 2011; Mitterer & Mattys, 2017). Note that most of these studies are conducted with repeated, temporally discrete speech stimuli. One recent study with continuous speech stimuli suggests that the processing of syllables was largely unaffected when attention is away from the auditory modality (Ding et al., 2018). Hence, it may be reasonable to argue that the processing of phonetic features, which constitute syllables, may remain unchanged when diverting attention away from the auditory modality. In line with this argument, the current study used two different analysis approaches and provided converging evidence that the neural processing of phonetic features in the continuous speech was not impacted when listeners divided attention away from the auditory modality (e.g., 3-back = active listening).

Note that for the TRF analysis, we observed lower EEG prediction accuracy for the 0-back condition than the other two conditions (active listening and 3-back). One plausible explanation may be that EEG from the 0-back condition was more contaminated by artifacts, e.g., muscular artifacts from responding to the visual targets. However, this may be unlikely the case because i) the number of actual responses was not different between 0- and 3-back conditions; and ii) the number of ICA components rejected due to ocular artifacts was not different between 0-back and the other two task

conditions. Future studies are needed to clarify the factors that contribute to the decrease in EEG prediction accuracy for the 0-back condition.

Our findings suggest that the processing of the suprasegmental speech features (amplitude envelope and F0) is more affected by attention relative to the processing of segmental features (phonetic features). In general, the representation of suprasegmental speech features requires information encoded in a longer temporal window than that of segmental features. According to the multi-time resolution processing model proposed by Hickok and Poeppel (2007), right-hemisphere mechanisms are predominant in extracting information over a longer temporal window, whereas integrating information over a short temporal window might rely on bilateral mechanisms. There is evidence to suggest a general right hemisphere dominance for attention (e.g., Asanowicz, Marzecová, Jaśkowski, & Wolski, 2012; Evert, McGlinchey-Berroth, Verfaellie, & Milberg, 2003), which might lead to more pronounced modulation of processing in the right hemisphere by attention. We speculate that this might be one plausible mechanism underlying stronger attentional modulation of envelope and F0 processing than phonetic processing observed in the current study, considering that the processing of envelope and F0 may be dominated in the right hemisphere. This speculative explanation is consistent with a recent intermodal attention study showing that attentional modulation of auditory processing was located mainly in the right auditory areas (Saupe, Schröger, Andersen, & Müller, 2009).

It should be noted that, when the demand of the visual task increased from 0-back to 3-back, we found a decrease in behavioral performance on the continuous speech stimuli but failed to observe a reduction in the neural processing of continuous speech stimuli. The divergence between behavioral and neural measures on the continuous speech stimuli was further evidenced by the findings that no significant correlations were

found between these two types of measures. Indeed, the relationship between the neural encoding of speech features and behavioral measures on the processing of continuous speech stimuli is debatable (Ding & Simon, 2014; Kösem & van Wassenhove, 2017). For example, regarding the processing of speech envelope, some studies found a positive link between the accuracy to represent speech envelope and behavioral speech comprehension performance on the corresponding continuous speech stimuli (Crosse et al., 2015; Ding & Simon, 2013; O'Sullivan et al., 2014), but other studies failed to observe such a link (Presacco et al., 2016; Puschmann et al., 2017). This may not be surprising considering that the behavioral speech tasks, i.e., answering questions related to the content of the continuous speech stimuli as in our and many other studies, involve a series of processes beyond merely encoding the acoustic envelope, F0, and phonetic features. These other processes include grouping phonemes into syllables, words, phrases, and sentences, and keeping the gist into working memory, many of which are likely to be affected by attention (Backer, Binns, & Alain, 2015; Ding et al., 2018; Lim, Wöstmann, & Obleser, 2015).

In summary, the present study demonstrated that, dividing attention between auditory and visual tasks leads to both behavioral and electrophysiological costs in the processing of continuous speech stimuli. Interestingly, our findings indicate that the neural encoding of suprasegmental features (e.g., envelope and fundamental frequency) in continuous speech can be modulated by diverting attention away from the auditory modality, whereas the neural encoding of segmental features (e.g., phonetic features) may be unaffected by taking attention away from the auditory stream.

GENERAL DISCUSSION

Across two studies, we investigated the extent to which taking attention away from the auditory modality to the visual modality influence the processing of speech signals. In general, findings from the two studies demonstrate behavioral and electrophysiological costs in the processing of speech signals when attention is diverted to visual tasks. These findings corroborate extensive evidence of visual attentional effects on auditory processing with temporally discrete non-speech sounds (e.g., Alho, Woods, Algazi, & Näätänen, 1992; Chait et al., 2012; Davis, 1964; Dyson, Alain, & He, 2005; Hackley, Woldorff, & Hillyard, 1990; Haroush, Hochstein, & Deouell, 2010; Karns & Knight, 2009; Molloy et al., 2015; Porcu et al., 2014; Woods, Alho, & Algazi, 1992; Zhang, Chen, Yuan, Zhang, & He, 2006) as well as continuous non-speech sounds (e.g., Keitel, Schröger, Saupe, & Müller, 2011; Saupe, Schröger, et al., 2009; Saupe, Widmann, Bendixen, Müller, & Schröger, 2009). Meanwhile, there is numerous evidence that attending to the auditory modality can exert detrimental effects on visual processing (e.g., Alho et al., 1992; Gherri & Eimer, 2011; Hackley et al., 1990; Karns & Knight, 2009; Murphy & Greene, 2017; Saupe, Schröger, et al., 2009; Sinnett et al., 2006; Woods et al., 1992). For example, Gherri and Eimer (2011) demonstrated that active listening to narrative stories decreases the perceptual processing of visual stimuli and the attentional selection of visual target. Together, these studies are in line with the *supramodal* account of attentional resources (Broadbent, 1957; Ciaramitaro, Chow, & Eglington, 2017; Jolicoeur, 1999; Klemen, Büchel, & Rose, 2009; Macdonald & Lavie, 2011; Molloy, Griffiths, Chait, & Lavie, 2015; Raveh & Lavie, 2015), but are inconsistent with the *modality-specific* account of attentional resources (Alais et al., 2006; Arrighi et al., 2011; Duncan et al., 1997; Keitel et al., 2013; Parks et al., 2011; Porcu et al., 2014).

Study 1 suggests that the impact of neural processing of speech signals from crossmodal attention may be as early as auditory subcortical levels. In this study, crossmodal attention was manipulated via varying the perceptual load of a visual search task. We found that visual perceptual load resulted in decreased early sensory representation of linguistically-relevant suprasegmental features (pitch patterns) when the auditory signals were presented in predictable contexts. This finding is consistent with the perceptual load theory that high perceptual load leads to an early selection of information, i.e., filtering of distractors at early processing stage (see for example Lavie, 2005; Murphy, Groeger, & Greene, 2016 for a review). Going beyond this finding, we further show that, when the speech signals were presented in variable (less predictable) contexts, increasing visual perceptual load was associated with enhanced early sensory representation of the linguistically-relevant pitch patterns. This is not what we would expect based on the perceptual load theory. Thus, study 1 provides novel evidence that the regularities of incoming stimuli may be a factor that can modulate the impact of perceptual load on sensory processing.

In study 2, we demonstrate that the processing of continuous speech signals is impaired by crossmodal attention. In this study, crossmodal attention was manipulated via varying the demands of a visuospatial working memory task. This finding is inconsistent with the perceptual load theory that increasing cognitive load leads to enhanced processing of distractors (see for example Lavie, 2010; Murphy et al., 2016 for a review), but is in keeping with the neurocognitive task-engagement/distraction trade-off model that increasing task difficulty, such as increasing cognitive (working memory) load, leads to enhanced level of active suppression of task-irrelevant stimuli (Sörqvist et al., 2016, 2012; Sörqvist & Marsh, 2015; Sörqvist & Rönnberg, 2014).

Findings from the two studies suggest that taking attention away from the auditory modality may modulate the processing of suprasegmental speech features (e.g., envelope and fundamental frequency). This knowledge may be used to infer whether a listener is attending to a speaker. This information, on one hand, would allow the speaker to make an adjustment to the spoken content to match the listener's interest, and on the other hand can be a neurofeedback signal to inform the listener to better focus on the speaker. This may be useful in many real-life scenarios, e.g., learning in classroom settings. Further, our findings suggest that disengaging attention from the auditory modality may not affect the processing of phonological information in continuous speech. Impairments in phonological processing has been proposed as a core deficit underlying developmental dyslexia (Ramus, 2003). Based on our finding, we may be able to develop more ecologically valid paradigms to assay phonological processing ability in children with developmental dyslexia (Di Liberto et al., 2018). In such paradigms, we would be less concerned that attentional factors, which may be associated with dyslexia (Stevens et al., 2013), confound the findings on phonological processing.

What is the neural basis for the crossmodal interaction between auditory and visual processing? There are at least two distinct but not necessarily exclusive mechanisms (Murray & Spierer, 2011). First, feedback projections from higher-order regions, e.g., regions of associative auditory cortex and frontal cortex, may mediate the interaction between auditory and visual processing (e.g., Durantin, Dehais, Gonthier, Terzibas, & Callan, 2017; Molloy et al., 2015; Yucel et al., 2005). In a recent fMRI study, Durantin et al. (2017) confronted participants with a visual perceptual-motor task while responding to audio alarms. They found that the functional connectivity between the inferior frontal gyrus and auditory cortex was reduced for missed audio alarms relative to successfully detected audio alarms. Second, the cortico-cortical connection

between auditory and visual cortices may drive the interaction between auditory and visual processing (e.g., Kayser, Logothetis, & Panzeri, 2010; Raij et al., 2010). Animal studies have documented anatomical connections between auditory and visual cortices (e.g., Budinger, Heil, Hess, & Scheich, 2006; Henschke, Noesselt, Scheich, & Budinger, 2015; Rockland & Van Hoesen, 1994; Stehberg, Dang, & Frostig, 2014). Functionally, Raij et al. (2010) suggest that the activation of primary auditory and visual cortices by stimuli from the opposite modality was about 10-55 ms later than that by the same-modality stimuli. Such delay is consistent with the conduction delays from one sensory cortex to another of 30-35 ms.

It is worth noting that the neural processing of speech signals was modulated by the demands of the visual task in study 1, but not in study 2. We argue that this may be due to that the task paradigm in study 1, relative to that in study 2, exerts stronger control of attention away from the speech stimuli. First, the paradigm in study 1 involves a single visual task while the speech stimuli were task-irrelevant. The paradigm in study 2 involves, however, a dual-task wherein participants need to respond not only a visual task but also questions on the speech stimuli. Hence, it is likely that participants would focus solely on the visual stimuli in study 1 but divide attention between the visual and auditory stimuli in study 2. Second, the speech stimuli in study 1 always overlapped with the visual stimuli that require speeded responses. In contrast, the speech stimuli in study 2 only partially overlap with the visual stimuli that not always require a response. The larger temporal overlap between auditory and visual stimuli in study 1 as opposed to study 2 possibly constrains to a larger extent the processing of the speech stimuli (Molloy et al., 2015; Pashler, 1994; Sigman & Dehaene, 2008).

IMPLICATIONS AND CONCLUSIONS

The present dissertation sought to examine the effects of taking attention away from the auditory modality on the processing of speech signals. Findings from two studies demonstrates that the neural encoding of suprasegmental features (e.g., envelope and F0) is subject to the influence of crossmodal attention, while the neural encoding of segmental features (e.g., phonetic features) may be unaffected by diverting attention away from the auditory stream. Moreover, study 1 suggests that the crossmodal attentional influence on speech processing may be as early as auditory subcortical levels. Critically, the impact of crossmodal attention on the early sensory encoding of speech signals is dependent on the predictability of the incoming speech stream, a possible reflection of the push-pull dynamic between predictive processes and novelty detection within the auditory system. Further, the two studies demonstrate the feasibility of machine learning approaches to assay early sensory representational changes as a function of biologically relevant influences, e.g., attention and stimulus statistics (study 1), and to evaluate different aspects of speech processing with ecologically valid stimuli (study 2). These findings may form the basis for future research on natural speech processing and for interventions on improving speech processing outcomes in real-world multisensory scenarios.

APPENDIX

Here we provide a list of the multiple-choice questions for the continuous speech stimuli in study 2. There are 45 tracks of continuous speech stimuli, with two questions for each track. The keys highlighted in yellow are the correct answers.

Track 1

1. Who is Alice with before she sees the white rabbit?
   1.) Her cat
   2.) Her friend Mabel
   3.) Her sister
   4.) Her mother
2. What color eyes did the white rabbit have?
   1.) Pink
   2.) Red
   3.) Blue
   4.) Green

Track 2

1. What does Alice pick up while falling through the well, a/an?
   1.) map
   2.) empty jar
   3.) bottle
   4.) book
2. What had been in the empty jar?
   1.) Orange jelly
   2.) Orange jam
   3.) Orange honey
   4.) Orange marmalade

Track 3

1. How many miles does Alice think she has fallen?
   1.) 2,000
   2.) 3,000
   3.) 4,000
   4.) 5,000
2. What does Alice try to do while talking to herself?
   1.) Curtsey
   2.) Bow
   3.) Shake hands

4.) High Five

Track 4

1. Who is Dinah?
   1.) Alice's dog
   2.) Alice's bat
   3.) Alice's mouse
   4.) Alice's cat
2. What does Alice want to ask Dinah?
   1.) Did you ever eat a rat?
   2.) Did you ever eat a bat?
   3.) Did you ever eat a squirrel?
   4.) Did you ever eat a mouse?

Track 5

1. Where does Alice find herself after falling down the hole?
   1.) A short hallway
   2.) A high hallway
   3.) A long hallway
   4.) A dark hallway
2. How many legs does the little table have?
   1.) Four
   2.) Five
   3.) Three
   4.) Two

Track 6

1. What does Alice see at the end of the passage, a?
   1.) Telescope
   2.) Garden
   3.) Wishing well
   4.) Rose bush
2. What does Alice find on top of the little table when she goes to it again?
   1.) A gold key
   2.) A little bottle
   3.) A big bottle
   4.) A book of rules

Track 7

1. What was written around the neck of the bottle?
   1.) 'DRINK ME'
   2.) 'DO NOT DRINK'
   3.) 'POISON'
   4.) 'TASTE ME'
2. Why didn't Alice drink from the bottle right away?
   1.) She wanted to find out if the bottle was marked 'poison'.

2.) She was not able to read the label.
3.) She wanted to find out who marked the bottle ‘drink me’.
4.) She wanted to share the potion with the rabbit.

Track 8

1. What happened to Alice after she finished the little bottle?
   1.) She felt very sick
   2.) She grew very large
   3.) She became very small
   4.) She played in the lovely garden
2. How does Alice try to reach the little golden key on the table?
   1.) Climb up one of the legs
   2.) Jump up to the table top
   3.) Drink from the little bottle
   4.) Knock the table down

Track 9

1. What does Alice find lying under the table? A little glass:
   1.) Bottle
   2.) Box
   3.) Slipper
   4.) jar
2. What happened to Alice when she ate a little bit of cake?
   1.) She grew larger
   2.) She grew smaller
   3.) She remained the same size
   4.) She began to cry

Track 10

1. What does Alice promise to give her feet every Christmas? A pair of:
   1.) stockings
   2.) boots
   3.) shoes
   4.) mittens
2. How tall does Alice become after eating the cake? More than:
   1.) 6 feet
   2.) 7 feet
   3.) 8 feet
   4.) 9 feet

Track 11

1. How deep was the pool of tears?
   1.) 3 inches
   2.) 4 inches
   3.) 5 inches
   4.) 6 inches

2. What was the White Rabbit carrying when Alice saw him? White kid gloves and a:
    1.) Waistcoat-pocket
    2.) fan
    3.) golden key
    4.) glass box

Track 12

1. What does Alice think happened to her the night before?
    1.) She had food poisoning
    2.) She had a nightmare
    3.) She was changed into a different person
    4.) She was changed into Ada
2. What does Alice start doing to test her past knowledge?
    1.) Multiplication problems
    2.) Addition problems
    3.) Subtraction problems
    4.) Division problems

Track 13

1. What reptile is in the lesson that Alice recites?
    1.) Lizard
    2.) Snake
    3.) Crocodile
    4.) Alligator
2. Who does Alice believe she has become?
    1.) Ada
    2.) Mabel
    3.) Dinah
    4.) Edie

Track 14

1. What made Alice shrink again?
    1.) The fan
    2.) The glass box
    3.) The little bottle
    4.) The white kid gloves
2. How tall was Alice when she measured herself?
    1.) 5 ft.
    2.) 4 ft.
    3.) 3 ft.
    4.) 2 ft.

Track 15

1. What does Alice fall in?
    1.) Mineral water
    2.) Spring water

3.) Salt water
4.) Lake water

2. Who does Alice see in the pool?
   1.) A walrus
   2.) A hippopotamus
   3.) A cat
   4.) A mouse

Track 16

1. At first, Alice thinks the mouse might speak:
   1.) Latin
   2.) French
   3.) Spanish
   4.) Portuguese
2. Alice wishes she could show the mouse her:
   1.) Cat
   2.) Dog
   3.) Bird
   4.) Rabbit

Track 17

1. Alice felt certain the mouse was:
   1.) delighted
   2.) disappointed
   3.) offended
   4.) relaxed
2. What kind of dog does Alice start talking about?
   1.) A terrier
   2.) A beagle
   3.) A retriever
   4.) A poodle

Track 18

1. Who led the way to the shore?
   1.) Mouse
   2.) Alice
   3.) Eaglet
   4.) Dodo
2. What was the first question asked on shore?
   1.) How to get food
   2.) How to get water
   3.) How to get the white rabbit
   4.) How to get dry again

Track 19

1. Who did Alice have a long argument with?
   1.) The Dodo

2.) The Eaglet
3.) The Lory
4.) The Duck

2. Who seemed to be the person of authority in the group?
   1.) The Eaglet
   2.) The Mouse
   3.) The Lory
   4.) The Dodo

Track 20

1. In mouse's story, Stigand was the patriotic archbishop of:
   1.) Cambridge
   2.) Durham
   3.) Canterbury
   4.) Salisbury
2. What was one of the things the Duck usually found?
   1.) Frogs
   2.) Snails
   3.) Salamanders
   4.) Weeds

Track 21

1. Who came up with the idea of a Caucus-Race?
   1.) The Eaglet
   2.) The Mouse
   3.) The Dodo
   4.) The Lory
2. The race course was marked out in a sort of:
   1.) Square
   2.) Triangle
   3.) Oval
   4.) Circle

Track 22

1. Who won the race?
   1.) The Dodo
   2.) The Lory
   3.) Everybody
   4.) Nobody
2. Who handed out prizes?
   1.) Alice
   2.) The Dodo
   3.) The Lory
   4.) The Duck

Track 23

1. What else was in Alice's pocket?

1.) A needle
2.) A thimble
3.) A button
4.) A piece of string

2. Who did the animals beg to tell them something more?
   1.) The Dodo
   2.) The Mouse
   3.) The Lory
   4.) Alice

Track 24

1. What does Alice do while the mouse is speaking? Come up with her own:
   1.) tale
   2.) puzzle
   3.) discovery
   4.) legend
2. Why does the mouse yell at Alice?
   1.) She is laughing
   2.) She is talking
   3.) She is not attending
   4.) She is not organizing

Track 25

1. Why does Alice want the mouse to come back?
   1.) Finish his story
   2.) Finish the race
   3.) Finish the knot
   4.) Finish the swim
2. Who does Alice wish was with her?
   1.) Ada
   2.) Mabel
   3.) Edie
   4.) Dinah

Track 26

1. What do all of the animals do?
   1.) Leave Alice
   2.) Finish the story
   3.) Go for a swim
   4.) Find Dinah
2. Who comes back to Alice?
   1.) The mouse
   2.) The rabbit
   3.) The Crab
   4.) The Canary

Track 27

1. Who is the White Rabbit worried about?
   1.) The Duchess
   2.) The Queen of Hearts
   3.) The Duke
   4.) The King of Hearts
2. Who does the White Rabbit think Alice is?
   1.) Julianna
   2.) Polly Ann
   3.) Mary Beth
   4.) Mary Ann

Track 28

1. What is engraved on the brass plate of the door?
   1.) 'W. RABBIT'
   2.) 'WHITE RABBIT'
   3.) 'WHITE R.'
   4.) 'MR. RABBIT'
2. What catches Alice's eye in the tidy little room?
   1.) a piece of cake
   2.) a little crumb
   3.) a little bottle
   4.) a big bottle

Track 29

1. What does Alice do to save her neck from being broken?
   1.) Swerve
   2.) straighten
   3.) droop
   4.) stoop
2. Where does Alice put one of her feet?
   1.) up the chimney
   2.) out the window
   3.) through the door
   4.) down the hallway

Track 30

1. What does Alice think should be written about her?
   1.) A book
   2.) A poem
   3.) A song
   4.) A lesson
2. What makes Alice stop to listen?
   1.) A crash
   2.) A cry
   3.) pattering of feet
   4.) A voice

Track 31

1. What does the Caterpillar take out of its mouth, a?
   1.) hookah
   2.) pipe
   3.) cigarette
   4.) lollipop
2. What does the Caterpillar ask Alice?
   1.) How did you get here?
   2.) What are you doing?
   3.) Where are you going?
   4.) Who are you?

Track 32

1. How does Alice feel after the Caterpillar's remarks?
   1.) Refreshed
   2.) Irritated
   3.) Delighted
   4.) Exasperated
2. Alice thinks the Caterpillar should tell her:
   1.) Who he is first
   2.) How to get home
   3.) Where to find the White Rabbit
   4.) How to become a butterfly

Track 33

1. What is the important thing that the Caterpillar has to say to Alice?
   1.) 'Keep your temper'
   2.) 'Keep calm'
   3.) 'Hold your tongue'
   4.) 'Move along'
2. What does the Caterpillar ask Alice to recite? You are Old, Father:
   1.) John
   2.) William
   3.) Edward
   4.) Abraham

Track 34

1. Alice had never been ___so much in her life.
   1.) charmed
   2.) delighted
   3.) contradicted
   4.) challenged
2. What was Alice's current height?
   1.) 2 inches
   2.) 3 inches
   3.) 4 inches

4.) 5 inches

Track 35

1. What does the Caterpillar get off?
   1.) mushroom
   2.) stool
   3.) chair
   4.) stump
2. What side of the mushroom does Alice eat first, the?
   1.) right side
   2.) left side
   3.) north side
   4.) south side

Track 36

1. What happened to Alice after she ate the right side of the mushroom?
   1.) She stayed the same size
   2.) She became taller
   3.) She became wider
   4.) She became smaller
2. What is the only thing Alice could see after she ate the other side of the mushroom? Her:
   1.) Feet
   2.) Shoulders
   3.) Hands
   4.) Neck

Track 37

1. Who does the Pigeon mistake Alice for?
   1.) A lizard
   2.) A serpent
   3.) A salamander
   4.) A frog
2. How long has it been since the Pigeon slept?
   1.) 5 weeks
   2.) 4 weeks
   3.) 3 weeks
   4.) 2 weeks

Track 38

1. What does the Pigeon accuse Alice of doing?
   1.) inventing something
   2.) discovering something
   3.) moving something
   4.) realizing something
2. What does the Pigeon suppose Alice will say next?
   1.) She has never eaten a Pigeon.

2.) She has never been a little girl.
3.) She has never tasted a mushroom.
4.) She has never tasted an egg.

Track 39

1. Why does the Pigeon feel threatened by Alice? She is afraid that Alice will:
   1.) grab her
   2.) eat her eggs
   3.) climb the tree
   4.) learn to fly
2. What does Alice remember she still has?
   1.) pieces of cake
   2.) the little bottle
   3.) pieces of mushroom
   4.) the glass box

Track 40

1. How tall does Alice make herself to go to the house?
   1.) 9 inches
   2.) 10 inches
   3.) 11 inches
   4.) 12 inches
2. What style hair did the footmen have?
   1.) crimped
   2.) flipped
   3.) braided
   4.) curled

Track 41

1. Which footman produces the great letter?
   1.) the frog footman
   2.) the fish footman
   3.) the dog footman
   4.) the queen footman
2. Why does the footman tell Alice that there is no use in knocking?
   1.) the noise within
   2.) no one is home
   3.) the dog within
   4.) everyone is playing croquet

Track 42

1. What was the footman doing all the time he was speaking?
   1.) looking for the Duchess
   2.) looking up into the sky
   3.) looking up into a tree
   4.) looking for the other footman
2. What grazes the footman's nose?

1.) a kettle
2.) a dish
3.) a plate
4.) a saucer

Track 43

1. What does Alice think is really dreadful about all the creatures?
   1.) The way they argue
   2.) The way they throw plates
   3.) The way they don't listen
   4.) The way they sit around all day
2. What is in the big cauldron in the kitchen?
   1.) Chili
   2.) Stew
   3.) Magic potion
   4.) Soup

Track 44

1. How does the Duchess address the baby?
   1.) 'Hog!'
   2.) 'Sow!'
   3.) 'Pig!'
   4.) 'Boar!'
2. What does the cook start throwing at the Duchess and the baby?
   1.) Nothing
   2.) Everything
   3.) Soup
   4.) Pepper

Track 45

1. What does the Duchess do when the sauce pans, plates, and dishes hit her?
   1.) She takes no notice to them
   2.) She yells "Chop off her head!"
   3.) She grabs her nose
   4.) She yells "That hurt!"
2. What is the cook doing instead of listening to the Duchess?
   1.) Throwing the soup
   2.) Stirring the tea
   3.) Shaking the soup
   4.) Stirring the soup

REFERENCES

Ahlfors, S. P., Han, J., Belliveau, J. W., & Hämäläinen, M. S. (2010). Sensitivity of MEG and EEG to source orientation. *Brain Topography*, *23*(3), 227–232.

Akhoun, I., Gallégo, S., Moulin, A., Ménard, M., Veuillet, E., Berger-Vachon, C., … Thai-Van, H. (2008). The temporal relationship between speech auditory brainstem responses and the acoustic pattern of the phoneme/ba/in normal-hearing adults. *Clinical Neurophysiology*, *119*(4), 922–933.

Alais, D., Morrone, C., & Burr, D. (2006). Separate attentional resources for vision and audition. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1592), 1339–1345. https://doi.org/10.1098/rspb.2005.3420

Alho, K., Woods, D. L., & Algazi, A. (1994). Processing of auditory stimuli during auditory and visual attention as revealed by event-related potentials. *Psychophysiology*, *31*(5), 469–479.

Alho, K., Woods, D. L., Algazi, A., & Näätänen, R. (1992). Intermodal selective attention. II.Effects of attentional load on processing of auditory and visual stimuli in central space. *Electroencephalography and Clinical Neurophysiology*, *82*, 356–368. https://doi.org/10.1016/0013-4694(92)90005-3

Anderson, L. A., & Malmierca, M. S. (2013). The effect of auditory cortex deactivation on stimulus-specific adaptation in the inferior colliculus of the rat. *European Journal of Neuroscience*, *37*(1), 52–62.

Arrighi, R., Lunardi, R., & Burr, D. (2011). Vision and audition do not share attentional resources in sustained tasks. *Frontiers in Psychology*, *2*(APR), 10–13. https://doi.org/10.3389/fpsyg.2011.00056

Asanowicz, D., Marzecová, A., Jaśkowski, P., & Wolski, P. (2012). Hemispheric asymmetry in the efficiency of attentional networks. *Brain and Cognition*, *79*(2), 117–128.

Ayala, Y. A., & Malmierca, M. S. (2013). Stimulus-specific adaptation and deviance detection in the inferior colliculus. *Frontiers in Neural Circuits*, *6*, 89.

Ayala, Y. A., Udeh, A., Dutta, K., Bishop, D., Malmierca, M. S., & Oliver, D. L. (2015).

Differences in the strength of cortical and brainstem inputs to SSA and non-SSA neurons in the inferior colliculus. *Scientific Reports*, *5*, 10383.

Backer, K. C., Binns, M. A., & Alain, C. (2015). Neural dynamics underlying attentional orienting to auditory representations in short-term memory. *Journal of Neuroscience*, *35*(3), 1307–1318.

Bajo, V. M., Nodal, F. R., Moore, D. R., & King, A. J. (2010). The descending corticocollicular pathway mediates learning-induced auditory plasticity. *Nature Neuroscience*, *13*(2), 253.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version*, *1*(7).

Batra, R., Kuwada, S., & Maher, V. L. (1986). The frequency-following response to continuous tones in humans. *Hearing Research*, *21*(2), 167–177.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bidelman, G. M. (2015a). Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR. *Hearing Research*, *323*, 68–80.

Bidelman, G. M. (2015b). Towards an optimal paradigm for simultaneously recording cortical and brainstem auditory evoked potentials. *Journal of Neuroscience Methods*, *241*, 94–100.

Bidelman, G. M., & Alain, C. (2015). Musical training orchestrates coordinated neuroplasticity in auditory brainstem and cortex to counteract age-related declines in categorical vowel perception. *Journal of Neuroscience*, *35*(3), 1240–1249.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences* (Vol. 17, pp. 97–110). Amsterdam.

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer (Version 5.3. 51)[Computer program]. *Amsterdam: Institute of Phonetic Sciences/University of Amsterdam. Recuperado de Http://www. Praat. Org*.

Bonte, M., Parviainen, T., Hytönen, K., & Salmelin, R. (2005). Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. *Cerebral Cortex*, *16*(1), 115–123.

Broadbent, D. E. (1952). Listening to one of two synchronous messages. *Journal of Experimental Psychology*, *44*(1), 51.

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*, *64*(3), 205.

Broadbent, D. E. (1958). *Perception and communication.* Pergamon Press.

Budinger, E., Heil, P., Hess, A., & Scheich, H. (2006). Multisensory processing via early cortical stages: connections of the primary auditory cortical field with other sensory systems. *Neuroscience*, *143*(4), 1065–1083.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001–1010.

Causse, M., Imbert, J.-P., Giraudet, L., Jouffrais, C., & Tremblay, S. (2016). The Role of Cognitive and Perceptual Loads in Inattentional Deafness. *Frontiers in Human Neuroscience*, *10*(July), 1–12. https://doi.org/10.3389/fnhum.2016.00344

Celesia, G. G., Broughton, R. J., Rasmussen, T., & Branch, C. (1968). Auditory evoked responses from the exposed human cortex. *Electroencephalography and Clinical Neurophysiology*, *24*(5), 458–465.

Chait, M., Ruff, C. C., Griffiths, T. D., & McAlpine, D. (2012). Cortical responses to changes in acoustic regularity are differentially modulated by attentional load. *NeuroImage*, *59*(2), 1932–1941. https://doi.org/10.1016/j.neuroimage.2011.09.006

Chandrasekaran, B., Hornickel, J., Skoe, E., Nicol, T., & Kraus, N. (2009). Context-dependent encoding in the human auditory brainstem relates to hearing speech in noise: implications for developmental dyslexia. *Neuron*, *64*(3), 311–319.

Chandrasekaran, B., & Kraus, N. (2010). The scalp-recorded brainstem response to speech: Neural origins and plasticity. *Psychophysiology*, *47*(2), 236–246.

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2007). Mismatch negativity to pitch contours is influenced by language experience. *Brain Research*, *1128*, 148–156.

Chandrasekaran, B., Skoe, E., & Kraus, N. (2014). An integrative model of subcortical auditory plasticity. *Brain Topography*, *27*(4), 539–552.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

Cherry, E. C., & Taylor, W. K. (1954). Some further experiments upon the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *26*(4), 554–559.

Ciaramitaro, V. M., Chow, H. M., & Eglington, L. G. (2017). Cross-modal attention influences auditory contrast sensitivity: Decreasing visual load improves auditory thresholds for amplitude- and frequency-modulated sounds. *Journal of Vision*, *17*(3), 20. https://doi.org/10.1167/17.3.20

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., & Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature Communications*, *7*, 11070. https://doi.org/10.1038/ncomms11070

Coffey, E. B. J., Mogilever, N., & Zatorre, R. J. (2017). Speech-in-noise perception in musicians: a review. *Hearing Research*.

Cohen, D., & Cuffin, B. N. (1983). Demonstration of useful differences between magnetoencephalogram and electroencephalogram. *Electroencephalography and Clinical Neurophysiology*, *56*(1), 38–51.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201.

Davis, H. (1964). Enhancement of Evoked Cortical Potentials in Humans Related to a Task Requiring a Decision. *Science*, *145*(3628), 182–183.

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80.

Di Liberto, G. M., & Lalor, E. C. (2017). Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. *Hearing Research*, *348*, 70–77.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465.

Di Liberto, G. M., Peter, V., Kalashnikova, M., Goswami, U., Burnham, D., & Lalor, E. C. (2018). Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia. *NeuroImage*.

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46. https://doi.org/10.1016/j.neuroimage.2013.10.054

Ding, N., Pan, X., Luo, C., Su, N., Zhang, W., & Zhang, J. (2018). Attention is required for knowledge-based sequential grouping: Insights from the integration of syllables into words. *Journal of Neuroscience*, *38*(5), 1178–1188.

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859.

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89.

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*(13), 5728–5735.

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, *87*(3), 272.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, *387*(6635), 808–810.

Duque, D., & Malmierca, M. S. (2015). Stimulus-specific adaptation in the inferior colliculus of the mouse: anesthesia and spontaneous activity effects. *Brain Structure and Function*, *220*(6), 3385–3398.

Durantin, G., Dehais, F., Gonthier, N., Terzibas, C., & Callan, D. E. (2017). Neural signature of inattentional deafness. *Human Brain Mapping*, *5455*, 5440–5455. https://doi.org/10.1002/hbm.23735

Dux, P. E., Ivanoff, J., Asplund, C. L., & Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved fMRI. *Neuron*, *52*(6), 1109–1120.

Dyson, B. j., Alain, C., & He, Y. (2005). Effects of visual attentional load on low-level auditory scene analysis. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(3), 319–338. https://doi.org/10.1097/00001756-200003200-00043

Evert, D. L., McGlinchey-Berroth, R., Verfaellie, M., & Milberg, W. P. (2003). Hemispheric asymmetries for selective attention apparent only with increased task demands in healthy participants. *Brain and Cognition*, *53*(1), 34–41.

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife*, *6*.

Fuglsang, S. A., Dau, T., & Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage*, *156*(April), 435–444.

https://doi.org/10.1016/j.neuroimage.2017.04.026

Galbraith, G. C., Olfman, D. M., & Huffman, T. M. (2003). Selective attention affects human brain stem frequency-following response. *Neuroreport*, *14*(5), 735–738.

Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*.

Gherri, E., & Eimer, M. (2011). Active listening impairs visual perception and selectivity: an ERP study of auditory dual-task costs on visual attention. *Journal of Cognitive Neuroscience*, *23*(4), 832–844. https://doi.org/10.1162/jocn.2010.21468

Goldenholz, D. M., Ahlfors, S. P., Hämäläinen, M. S., Sharon, D., Ishitobi, M., Vaina, L. M., & Stufflebeam, S. M. (2009). Mapping the signal-to-noise-ratios of cortical sources in magnetoencephalography and electroencephalography. *Human Brain Mapping*, *30*(4), 1077–1086.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." *Journal of Neuroscience*, *33*(4), 1417–1426.

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, *39*(3), 192–193.

Gósy, M., & Terken, J. M. B. (1994). Question marking in Hungarian: timing and height of pitch peaks.

Hackley, S. A., Woldorff, M., & Hillyard, S. A. (1990). Cross-Modal Selective Attention Effects on Retinal, Myogenic, Brainstem, and Cerebral Evoked Potentials. *Psychophysiology*, *27*(2), 195–208. https://doi.org/10.1111/j.1469-8986.1990.tb00370.x

Hadar, B., Skrzypek, J. E., Wingfield, A., & Ben-David, B. M. (2016). Working memory load affects processing time in spoken word recognition: Evidence from eye-movements. *Frontiers in Neuroscience*, *10*(MAY), 1–10. https://doi.org/10.3389/fnins.2016.00221

Hairston, W. D., Letowski, T. R., & McDowell, K. (2013). Task-Related Suppression of the Brainstem Frequency following Response. *PLoS ONE*, *8*(2), 31–34. https://doi.org/10.1371/journal.pone.0055215

Halin, N., Marsh, J. E., & Sörqvist, P. (2015). Central load reduces peripheral processing: Evidence from incidental memory of background speech. *Scandinavian Journal of Psychology*, *56*(6), 607–612. https://doi.org/10.1111/sjop.12246

Haroush, K., Hochstein, S., & Deouell, L. Y. (2010). Momentary fluctuations in allocation of attention: cross-modal effects of visual task load on auditory discrimination. *Journal of Cognitive Neuroscience*, *22*(7), 1440–51. https://doi.org/10.1162/jocn.2009.21284

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, *87*, 96–110.

Henschke, J. U., Noesselt, T., Scheich, H., & Budinger, E. (2015). Possible anatomical pathways for short-latency multisensory integration processes in primary sensory cortices. *Brain Structure and Function*, *220*(2), 955–977.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393.

Holmes, E., Purcell, D. W., Carlyon, R. P., Gockel, H. E., & Johnsrude, I. S. (2018). Attentional Modulation of Envelope-Following Responses at Lower (93–109 Hz) but Not Higher (217–233 Hz) Modulation Rates. *Journal of the Association for Research in Otolaryngology*, *19*(1), 83–97.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones* (Vol. 6). Cambridge University Press.

Hunter, C. R., & Pisoni, D. B. (2017). Extrinsic Cognitive Load Impairs Spoken Word Recognition in High- and Low-Predictability Sentences. *Ear and Hearing*, 1. https://doi.org/10.1097/AUD.0000000000000493

Jaeggi, S. M., Buschkuehl, M., Etienne, A., Ozdoba, C., Perrig, W. J., & Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(2), 75–89.

Johnson, K. L., Nicol, T. G., & Kraus, N. (2005). Brain stem response to speech: a biological marker of auditory processing. *Ear and Hearing*, *26*(5), 424–434.

Jolicoeur, P. (1999). Restricted attentional capacity between sensory modalities. *Psychonomic Bulletin and Review*, *6*(1), 87–92. https://doi.org/10.3758/BF03210813

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Prentice-Hall Englewood Cliffs, NJ.

Karns, C. M., & Knight, R. T. (2009). Intermodal auditory, visual, and tactile attention modulates early stages of neural processing. *Journal of Cognitive Neuroscience*, *21*(4), 669–83. https://doi.org/10.1162/jocn.2009.21037

Kastner, S., & Ungerleider, L. G. (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologia*, *39*(12), 1263–1276.

Kayser, C., Logothetis, N. K., & Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. *Current Biology*, *20*(1), 19–24.

Keitel, C., Maess, B., Schröger, E., & Müller, M. M. (2013). Early visual and auditory processing rely on modality-specific attentional resources. *NeuroImage*, *70*, 240–249. Retrieved from https://doi.org/10.1016/j.neuroimage.2012.12.046

Keitel, C., Schröger, E., Saupe, K., & Müller, M. M. (2011). Sustained selective intermodal attention modulates processing of language-like stimuli. *Experimental Brain Research*, *213*(2–3), 321–327. https://doi.org/10.1007/s00221-011-2667-2

Khalighinejad, B., Cruzatto da Silva, G., & Mesgarani, N. (2017). Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *The Journal of Neuroscience*, *37*(8), 2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017

King, A., Hopkins, K., & Plack, C. J. (2016). Differential group delay of the frequency following response measured vertically and horizontally. *Journal of the Association for Research in Otolaryngology*, *17*(2), 133–143.

King, A. J., & Bajo, V. M. (2013). Cortical modulation of auditory processing in the midbrain. *Frontiers in Neural Circuits*, *6*, 114.

Klemen, J., Büchel, C., & Rose, M. (2009). Perceptual load interacts with stimulus

processing across sensory modalities. *European Journal of Neuroscience*, *29*(12), 2426–2434. https://doi.org/10.1111/j.1460-9568.2009.06774.x

Knerr, S., Personnaz, L., & Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing* (pp. 41–50). Springer.

Kong, Y.-Y., Mullangi, A., & Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research*, *316*, 73–81.

Kösem, A., & van Wassenhove, V. (2017). Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, *32*(5), 536–544. https://doi.org/10.1080/23273798.2016.1238495

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*, 3–24.

Kraus, N., & White-Schwoch, T. (2015). Unraveling the biology of auditory learning: a cognitive–sensorimotor–reward framework. *Trends in Cognitive Sciences*, *19*(11), 642–654.

Kreitz, C., Furley, P., Simons, D. J., & Memmert, D. (2016). Does working memory capacity predict cross-modally induced failures of awareness? *Consciousness and Cognition*, *39*, 18–27. https://doi.org/10.1016/j.concog.2015.11.010

Krishnan, A. (2002). Human frequency-following responses: representation of steady-state synthetic vowels. *Hearing Research*, *166*(1–2), 192–201.

Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, *25*(1), 161–168.

Krishnan, A., Xu, Y., Gandour, J. T., & Cariani, P. A. (2004). Human frequency-following response: representation of pitch contours in Chinese tones. *Hearing Research*, *189*(1–2), 1–12.

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can

be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*(1), 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology*, *102*(1), 349–359.

Lau, J. C. Y., Wong, P. C. M., & Chandrasekaran, B. (2016). Context-dependent plasticity in the subcortical encoding of linguistic pitch patterns. *Journal of Neurophysiology*, *117*(2), 594–603.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 451.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, *9*(2), 75–82.

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science*, *19*(3), 143–148.

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, *56*(2), 183–197. https://doi.org/10.3758/BF03213897

Lehmann, A., Arias, D. J., & Schönwiesner, M. (2016). Tracing the neural basis of auditory entrainment. *Neuroscience*, *337*, 306–314.

Lim, S.-J., Wöstmann, M., & Obleser, J. (2015). Selective attention to auditory memory neurally enhances perceptual precision. *Journal of Neuroscience*, *35*(49), 16094–16104.

Llanos, F., Xie, Z., & Chandrasekaran, B. (2017). Hidden Markov modeling of frequency-following responses to Mandarin lexical tones. *Journal of Neuroscience Methods*, *291*, 101–112.

Macdonald, J. S. P., & Lavie, N. (2011). Visual perceptual load induces inattentional deafness. *Attention, Perception, & Psychophysics*, *73*(6), 1780–1789. https://doi.org/10.3758/s13414-011-0144-4

Makov, S., Sharon, O., Ding, N., Ben-Shachar, M., Nir, Y., & Zion Golumbic, E. (2017).

Sleep Disrupts High-Level Speech Parsing Despite Significant Basic Auditory Processing. *The Journal of Neuroscience*, *37*(32), 7772–7781. https://doi.org/10.1523/JNEUROSCI.0168-17.2017

Malmierca, M. S., Anderson, L. A., & Antunes, F. M. (2015). The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: a potential neuronal correlate for predictive coding. *Frontiers in Systems Neuroscience*, *9*, 19.

Malmierca, M. S., Cristaudo, S., Pérez-González, D., & Covey, E. (2009). Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *Journal of Neuroscience*, *29*(17), 5483–5493.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, *9*(6), 296–305.

Marsh, J. T., Worden, F. G., & Smith, J. C. (1970). Auditory frequency-following response: neural or artifact? *Science*, *169*(3951), 1222–1223.

Masutomi, K., Barascud, N., Kashino, M., McDermott, J. H., & Chait, M. (2016). Sound segregation via embedded repetition is robust to inattention. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 386–400. https://doi.org/10.1037/xhp0000147

Mattys, S. L., Barden, K., & Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin & Review*, *21*(3), 748–54. https://doi.org/10.3758/s13423-013-0544-7

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*(2), 145–160. https://doi.org/10.1016/j.jml.2011.04.004

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 1245994.

Mirkovic, B., Debener, S., Jaeger, M., & De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life

applications. *Journal of Neural Engineering*, *12*(4), 46007.

Mitterer, H., & Mattys, S. L. (2017). How does cognitive load influence speech perception? An encoding hypothesis. *Attention, Perception, & Psychophysics*, *79*(1), 344–351. https://doi.org/10.3758/s13414-016-1195-3

Møller, A. R., & Jannetta, P. J. (1982). Evoked potentials from the inferior colliculus in man. *Electroencephalography and Clinical Neurophysiology*, *53*(6), 612–620.

Møller, A. R., & Jannetta, P. J. (1985). Neural generators of the auditory brainstem response. *The Auditory Brainstem Response*, 13–31.

Molloy, K., Griffiths, T. D., Chait, M., & Lavie, N. (2015). Inattentional deafness: visual load leads to time-specific suppression of auditory evoked responses. *Journal of Neuroscience*, *35*(49), 16046–16054.

Moushegian, G., Rupert, A. L., & Stillman, R. D. (1973). Scalp-recorded early responses in man to frequencies in the speech range. *Electroencephalography and Clinical Neurophysiology*, *35*(6), 665–667.

Murphy, G., & Greene, C. M. (2017). The Elephant in the Road: Auditory Perceptual Load Affects Driver Perception and Awareness. *Applied Cognitive Psychology*, *31*(2), 258–263. https://doi.org/10.1002/acp.3311

Murphy, G., Groeger, J. A., & Greene, C. M. (2016). Twenty years of load theory—Where are we now, and where should we go next? *Psychonomic Bulletin & Review*, *23*(5), 1316–1340.

Murray, M. M., & Spierer, L. (2011). Multisensory integration: what you see is where you hear. *Current Biology*, *21*(6), R229–R231.

Musacchia, G., Strait, D., & Kraus, N. (2008). Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. *Hearing Research*, *241*(1–2), 34–42.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, *24*(4), 375–425.

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified

events. *Cognitive Psychology*, *7*(4), 480–494.

Nelken, I., & Ulanovsky, N. (2007). Mismatch negativity and stimulus-specific adaptation in animal models. *Journal of Psychophysiology*, *21*(3–4), 214–223.

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2017). Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading. *Frontiers in Human Neuroscience*, *10*(January), 679. https://doi.org/10.3389/fnhum.2016.00679

O'sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706.

Oatman, L. C., & Anderson, B. W. (1977). Effects of visual attention on tone burst evoked auditory potentials. *Experimental Neurology*, *57*(1), 200–211.

Oatman, L. C., & Anderson, B. W. (1980). Suppression of the auditory frequency following response during visual attention. *Electroencephalography and Clinical Neurophysiology*, *49*(3), 314–322.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*(4), 713–719.

Parbery-Clark, A., Strait, D. L., & Kraus, N. (2011). Context-dependent encoding in the auditory brainstem subserves enhanced speech-in-noise perception in musicians. *Neuropsychologia*, *49*(12), 3338–3345.

Parks, N. A., Hilimire, M. R., & Corballis, P. M. (2011). Steady-state signatures of visual perceptual load, multimodal distractor filtering, and neural competition. *Journal of Cognitive Neuroscience*, *23*(5), 1113–24. https://doi.org/10.1162/jocn.2010.21460

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological Bulletin*, *116*(2), 220.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*

Learning Research, *12*(Oct), 2825–2830.

Pérez-González, D., Malmierca, M. S., & Covey, E. (2005). Novelty detector neurons in the mammalian auditory midbrain. *European Journal of Neuroscience*, *22*(11), 2879–2885.

Phipson, B., & Smyth, G. K. (2010). Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, *9*(1).

Porcu, E., Keitel, C., & Müller, M. M. (2014). Visual, auditory and tactile stimuli compete for early sensory processing capacities within but not between senses. *NeuroImage*, *97*(May), 224–235. https://doi.org/10.1016/j.neuroimage.2014.04.024

Power, A. J., Foxe, J. J., Forde, E., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*(9), 1497–1503.

Presacco, A., Simon, J. Z., & Anderson, S. (2016). Effect of informational content of noise on speech representation in the aging midbrain and cortex. *Journal of Neurophysiology*, *116*(5), 2356–2367. https://doi.org/10.1152/jn.00373.2016

Puschmann, S., Steinkamp, S., Gillich, I., Mirkovic, B., Debener, S., & Thiel, C. M. (2017). The right temporoparietal junction supports speech tracking during selective listening: Evidence from concurrent EEG-fMRI. *Journal of Neuroscience*, 1007–1017.

Puvvada, K. C., & Simon, J. Z. (2017). Cortical Representations of Speech in a Multi-talker Auditory Scene. *Journal of Neuroscience*, *37*(Xx), 1–8. https://doi.org/10.1523/JNEUROSCI.0938-17.2017

Raij, T., Ahveninen, J., Lin, F., Witzel, T., Jääskeläinen, I. P., Letham, B., … Stufflebeam, S. (2010). Onset timing of cross-sensory activations and multisensory interactions in auditory and visual sensory cortices. *European Journal of Neuroscience*, *31*(10), 1772–1782.

Ramus, F. (2003). Developmental dyslexia: specific phonological deficit or general sensorimotor dysfunction? *Current Opinion in Neurobiology*, *13*(2), 212–218.

Raveh, D., & Lavie, N. (2015). Load-induced inattentional deafness. *Attention, Perception, & Psychophysics*, *77*(2), 483–492. https://doi.org/10.3758/s13414-014-0776-2

Rees, G., Frith, C., & Lavie, N. (2001). Processing of irrelevant cisual motion during performance of an auditory attention task. *Neuropsychologia*, *39*, 937–949.

Rockland, K. S., & Van Hoesen, G. W. (1994). Direct temporal-occipital feedback connections to striate cortex (V1) in the macaque monkey. *Cerebral Cortex*, *4*(3), 300–313.

Rodero, E. (2011). Intonation and emotion: influence of pitch levels and contour type on creating emotions. *Journal of Voice*, *25*(1), e25–e34.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *336*(1278), 367–373.

Sadeghian, A., Dajani, H. R., & Chan, A. D. C. (2015). Classification of speech-evoked brainstem responses to English vowels. *Speech Communication*, *68*, 69–84.

Saupe, K., Schröger, E., Andersen, S. K., & Müller, M. M. (2009). Neural mechanisms of intermodal sustained selective attention with concurrently presented auditory and visual stimuli. *Frontiers in Human Neuroscience*, *3*(November), 1–13. https://doi.org/10.3389/neuro.09.058.2009

Saupe, K., Widmann, A., Bendixen, A., Müller, M. M., & Schröger, E. (2009). Effects of intermodal attention on the auditory steady-state response and the event-related potential. *Psychophysiology*, *46*(2), 321–327. https://doi.org/10.1111/j.1469-8986.2008.00765.x

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime: User's guide*. Psychology Software Incorporated.

Shannon, R. V, Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303.

Shiga, T., Althen, H., Cornella, M., Zarnowiec, K., Yabe, H., & Escera, C. (2015). Deviance-related responses along the auditory hierarchy: Combined FFR, MLR and

MMN evidence. *PloS One*, *10*(9), e0136794.

Shomstein, S., & Yantis, S. (2004). Control of attention shifts between vision and audition in human cortex. *Journal of Neuroscience*, *24*(47), 10702–10706.

Sigman, M., & Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during dual-task performance. *Journal of Neuroscience*, *28*(30), 7585–7598.

Simon, S. S., Tusch, E. S., Holcomb, P. J., & Daffner, K. R. (2016). Increasing Working Memory Load Reduces Processing of Cross-Modal Task-Irrelevant Stimuli Even after Controlling for Task Difficulty and Executive Capacity. *Frontiers in Human Neuroscience*, *10*(August), 380. https://doi.org/10.3389/fnhum.2016.00380

Sinnett, S., Costa, A., & Soto-Faraco, S. (2006). Manipulating inattentional blindness within and across sensory modalities. *Quarterly Journal of Experimental Psychology*, *59*(8), 1425–1442. https://doi.org/10.1080/17470210500298948

Skoe, E., & Kraus, N. (2010). Auditory brainstem response to complex sounds: a tutorial. *Ear and Hearing*, *31*(3), 302.

Skoe, E., Krizman, J., Anderson, S., & Kraus, N. (2013). Stability and plasticity of auditory brainstem function across the lifespan. *Cerebral Cortex*, *25*(6), 1415–1426.

Slabu, L., Grimm, S., & Escera, C. (2012). Novelty detection in the human auditory brainstem. *Journal of Neuroscience*, *32*(4), 1447–1452.

Slee, S. J., & David, S. V. (2015). Rapid task-related plasticity of spectrotemporal receptive fields in the auditory midbrain. *Journal of Neuroscience*, *35*(38), 13090–13102.

Smith, J. C., Marsh, J. T., & Brown, W. S. (1975). Far-field recorded frequency-following responses: evidence for the locus of brainstem sources. *Clinical Neurophysiology*, *39*(5), 465–472.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34.

Sörqvist, P., Dahlström, Ö., Karlsson, T., & Rönnberg, J. (2016). Concentration: The neural underpinnings of how cognitive load shields against distraction. *Frontiers in*

Human Neuroscience, *10*, 221.

Sörqvist, P., & Marsh, J. E. (2015). How Concentration Shields Against Distraction. *Current Directions in Psychological Science*, *24*(4), 267–272. https://doi.org/10.1177/0963721415577356

Sörqvist, P., & Rönnberg, J. (2014). Individual differences in distractibility: An update and a model. *PsyCh Journal*, *3*(1), 42–57. https://doi.org/10.1002/pchj.47

Sörqvist, P., Stenfelt, S., & Rönnberg, J. (2012). Working Memory Capacity and Visual–Verbal Cognitive Load Modulate Auditory–Sensory Gating in the Brainstem: Toward a Unified View of Attention. *Journal of Cognitive Neuroscience*, *24*(11), 2147–2154. https://doi.org/10.1162/jocn_a_00275

Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K., & Chait, M. (2017). Is predictability salient? A study of attentional capture by auditory patterns. *Phil. Trans. R. Soc. B*, *372*(1714), 20160105.

Spence, C. (2010). Crossmodal attention. *Scholarpedia*, *5*(5), 6309.

Srinivasan, S., Keil, A., Stratis, K., Carr, K. L. W., & Smith, D. W. (2012). Effects of cross-modal selective attention on the sensory periphery: cochlear sensitivity is altered by selective attention. *Neuroscience*, *223*, 325–332.

Stefanics, G., Kremláček, J., & Czigler, I. (2014). Visual mismatch negativity: a predictive coding view. *Frontiers in Human Neuroscience*, *8*, 666.

Stehberg, J., Dang, P. T., & Frostig, R. D. (2014). Unimodal primary sensory cortices are directly connected by long-range horizontal projections in the rat sensory cortex. *Frontiers in Neuroanatomy*, *8*, 93.

Stevens, C., Harn, B., Chard, D. J., Currin, J., Parisi, D., & Neville, H. (2013). Examining the role of attention and instruction in at-risk kindergarteners: Electrophysiological measures of selective auditory attention before and after an early literacy intervention. *Journal of Learning Disabilities*, *46*(1), 73–86.

Suga, N. (2008). Role of corticofugal feedback in hearing. *Journal of Comparative Physiology A*, *194*(2), 169–183.

Suga, N., Yan, J., & Zhang, Y. (1997). Cortical maps for hearing and egocentric selection

for self-organization. *Trends in Cognitive Sciences*, *1*(1), 13–20.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215.

Sussman, E. S. (2013). Attention matters: pitch vs. pattern processing in adolescence. *Frontiers in Psychology*, *4*, 333.

Tang, C., Hamilton, L. S., & Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science*, *357*(6353), 797–801.

Team, R. C. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016.

Tobimatsu, S., & Celesia, G. G. (2006). Studies of human visual pathophysiology with visual evoked potentials. *Clinical Neurophysiology*, *117*(7), 1414–1433.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), 25–34. https://doi.org/10.1016/j.cognition.2005.01.006

Torralbo, A., Kelley, T. A., Rees, G., & Lavie, N. (2016). Attention induced neural response trade-off in retinotopic cortex under load. *Scientific Reports*, *6*, 33041.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review*, *76*(3), 282.

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, & Psychophysics*, *79*(2), 396–403.

van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Frontiers in Psychology*, *4*, 388.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181–1186.

Varghese, L., Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2015). Evidence against attentional state modulating scalp-recorded auditory brainstem steady-state responses. *Brain Research*, *1626*, 146–164.

https://doi.org/10.1016/j.brainres.2015.06.038

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology*, *37*(2), 190–203.

Wickens, C. D. (1991). Processing resources and attention. *Multiple-Task Performance*, *1991*, 3–34.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455.

Winer, J. A. (2005). Decoding the auditory corticofugal systems. *Hearing Research*, *207*(1–2), 1–9.

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, *13*(12), 532–540.

Wittekindt, A., Kaiser, J., & Abel, C. (2014). Attentional modulation of the inner ear: a combined otoacoustic emission and EEG study. *Journal of Neuroscience*, *34*(30), 9995–10002.

Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., Pantev, C., Sobel, D., & Bloom, F. E. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proceedings of the National Academy of Sciences*, *90*(18), 8722–8726.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*(4), 420.

Woods, D. L., Alho, K., & Algazi, A. (1992). Intermodal selective attention. I. Effects on event-related potentials to lateralized auditory and visual stimuli. *Electroencephalography and Clinical Neurophysiology*, *82*(5), 341–355. https://doi.org/10.1016/0013-4694(92)90004-2

Worden, F. G., & Marsh, J. T. (1968). Frequency-following (microphonic-like) neural responses evoked by sound. *Electroencephalography and Clinical Neurophysiology*, *25*(1), 42–52.

Xi, J., Zhang, L., Shu, H., Zhang, Y., & Li, P. (2010). Categorical perception of lexical tones in Chinese revealed by mismatch negativity. *Neuroscience*, *170*(1), 223–231.

Xie, Z., Reetzke, R., & Chandrasekaran, B. (2017). Stability and plasticity in neural encoding of linguistically relevant pitch patterns. *Journal of Neurophysiology*, *117*(3), 1407–1422.

Xie, Z., Yi, H.-G., & Chandrasekaran, B. (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PloS One*, *9*(12), e114439.

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. Laboratoire Parole et Langage, France.

Yan, J., & Suga, N. (1996). Corticofugal modulation of time-domain processing of biosonar information in bats. *Science*, *273*(5278), 1100–1103.

Yi, H. G., Xie, Z., Reetzke, R., Dimakis, A. G., & Chandrasekaran, B. (2017). Vowel decoding from single-trial speech-evoked electrophysiological responses: A feature-based machine learning approach. *Brain and Behavior*, *7*(6).

Yucel, G., Petty, C., McCarthy, G., & Belger, A. (2005). Graded visual attention modulates brain responses evoked by task-irrelevant auditory pitch changes. *Journal of Cognitive Neuroscience*, *17*(12), 1819–1828.

Zhang, P., Chen, X., Yuan, P., Zhang, D., & He, S. (2006). The effect of visuospatial attentional load on the processing of irrelevant acoustic distractors. *NeuroImage*, *33*(2), 715–724. https://doi.org/10.1016/j.neuroimage.2006.07.015

Zhang, Y., & Suga, N. (1997). Corticofugal amplification of subcortical responses to single tone stimuli in the mustached bat. *Journal of Neurophysiology*, *78*(6), 3489–3492.